Tomasz Magdziarz

Sektorowy formalizm porównawczej analizy powierzchni cząsteczkowej (s-CoMSA) – zastosowanie do modelowania zależności struktura-aktywność

Promotor pracy:
prof. dr hab. Jarosław Polański

Uniwersytet Śląski
Instytut Chemii
Katowice 2007

„A droga wiedzie w przód i w przód..."

J.R.R. Tolkien

Składam serdeczne podziękowania
Panu Prof. dr hab. Jarosławowi Polańskiemu
za wszelką pomoc, zrozumienie i opiekę naukową

Spis treści

Notacja

Notacja używana w pracy:

- a – skalar
  małe litery, kursywa
- $\boldsymbol{a}$ – wektor
  małe litery, pogrubione, kursywa
- a_i – element wektora
  małe litery, kursywa, jeden indeks
- $\boldsymbol{A}$ – macierz
  duże litery, pogrubione, kursywa
- a_{ij} – element macierzy
  małe litery, kursywa, dwa indeksy
- f( ) – funkcje
  małe litery, nawiasy
- **a10** – numery związków
  litera, liczba, pogrubione
- **atom.xyz** – nazwy pól struktur danych
  małe litery, pogrubione
- `prepro` – wartości pól struktur danych
  pismo maszynowe
- `Field*Coeff` – nazwy opcji programów komputerowych
  pismo maszynowe
- `qdb.mat` – nazwy plików, wartości pól struktur danych
  pismo maszynowe

Wykaz skrótów

Najważniejsze skróty używane w tekście pracy:

- 3D-QSAR – Three Dimensional QSAR
- CoMFA – Comparative Molecular Field Analysis
- CoMSA – Comparative Molecular Surface Analysis
- CBG – Corticosteroid-Binding Globulin
- cs – Cell Size
- CV – Cross-Validation
- IVE – Iterative Variable Elimination
- LOO – Live One Out
- LSO – Live Several Out
- PCA – Principal Components Analysis
- PLS – Partial Least Squares
- QSAR – Quantitative Structure – Activity Relationships
- RMS – Root Mean Square
- s-CoMSA – Sector Comparative Molecular Surface Analysis
- SAR – Structure – Activity Relationships
- SDEP – Standard Deviation of Error of Prediction
- SOM – Seflf-Organizing Maps
- SOM-CoMSA – (SOM)-Comparative Molecular Surface Analysis
- UVE – Uninformative Variable elimination

1 Cel pracy

Celem pracy jest:

- Opracowanie nowej metody obliczania deskryptorów cząsteczkowych s-CoMSA (ang. sector-comparative molecular surface analysis); w metodzie tej przestrzeń cząsteczkowa jest dzielona za zbiór sześciennych sektorów,
- Szeroko rozumiana optymalizacja metody s-CoMSA,
- Analiza QSAR oraz SAR wybranych szeregów związków aktywnych biologicznie z wykorzystaniem metody s-CoMSA oraz innych metod 3D-QSAR.

W zakres pracy wchodzi:

- Opracowanie formalizmu metody s-CoMSA,
- Zaprogramowanie procedur analizy s-CoMSA,
- Badanie modeli s-CoMSA aktywności biologicznej wybranych szeregów związków organicznych, w tym:
  - szeregu steroidów o powinowactwie do globuliny wiążącej kortykosteroidy (ang. corticosteroid-binding globulin – CBG),
  - inhibitorów wirusa HIV,
  - inhibitorów reduktazy kwasu dihydrofoliowego.

2 Wybrane metody modelowania wielowymiarowych zależności QSAR (m-QSAR)

Metody 3D-QSAR stanowią jedną z wielu prób racjonalizacji metod projektowania i poszukiwania leków. Podstawą metod QSAR oraz SAR (ang. Quantitative Structure-Activity Relationships – QSAR; ang. Structure-Activity Relationships – SAR) jest założenie, że aktywność cząsteczek jest funkcją ich struktury. W roku 1886 Crum-Brown i Fraser opublikowali pracę opisującą badania nad wpływem alkilowania atomów azotu w alkaloidach na ich działanie narkotyczne [1, 2]. Zaobserwowaną zależność wyrazili wzorem:

$$\Phi = f(C) \tag{2.1}$$

gdzie Φ oznacza działanie narkotyczne a C strukturę chemiczną. Podobne relacje zaobserwowali Richet, Meyer i Overton [3, 4, 5]. Podstawy klasycznych metod QSAR stworzyli niezależnie od siebie Hansch i Fujita oraz Free i Wilson [6, 7].

Powiązanie aktywności ze strukturą przez precyzyjne funkcje matematyczne nie jest łatwe. Struktura związku musi być przedstawiona w postaci numerycznej za pomocą tzw. deskryptorów. W klasycznych metodach QSAR wykorzystuje się deskryptory obliczone na podstawie dwuwymiarowej struktury cząsteczek. Cząsteczki nie są jednak obiektami dwuwymiarowymi. Tak więc obok metod 2D rozwinęły się metody wielowymiarowe uwzględniające rzeczywistą trójwymiarową strukturę cząsteczki, jej zmienność konformacyjną, sposób oddziaływania z receptorem a nawet z cząsteczkami rozpuszczalnika.

2.1 Metody 3D-QSAR

Rozwinięciem klasycznych metod QSAR jest zastosowanie takich deskryptorów, które są obliczane na podstawie trójwymiarowej struktury cząsteczki. Metody wykorzystujące deskryptory obliczone dla trójwymiarowych statycznych obrazów cząsteczek bez uwzględnienia ich labilności konformacyjnej nazywane są metodami 3D-QSAR.

2.1.1 Metoda CoMFA

Pierwszą metodą QSAR uwzględniającą trójwymiarową budowę cząsteczek była metoda DYLOMMS (ang. dynamic lattice-oriented molecular modelling system) [8, 9]. Polega ona na porównywaniu nałożonych na siebie cząsteczek chemicznych odwzorowanych na tzw. pola molekularne. Pole molekularne tworzy w tej metodzie trójwymiarowa sieć utworzona przez regularny układ sześcianów. Wartości potencjałów policzone w węzłach sieci tworzą deskryptory, które używane są do modelowania zależności QSAR. Ze względu na problemy obliczeniowe (porównaj rozdział 3.2, strona 25) metody DYLOMMS nie udało się zastosować praktycznie. W 1988 Cramer opisał metodę porównawczej analizy pól cząsteczkowych CoMFA (ang. comparative molecular field analysis) [10]. Metoda ta jest rozwinięciem metody DYLOMMS, która uzupełniona została o matematyczny aparat analizy PLS (porównaj rozdział 3.4, strona 30).

CoMFA mimo upływu czasu nadal jest cennym narzędziem modelowania 3D-QSAR [11, 12]. Często korzysta się z niej także w przypadku oceny i porównywania innych procedur modelowania QSAR. Swoją popularność metoda ta zawdzięcza pojawieniu się oprogramowania realizującego analizę CoMFA [13].

Modelowanie metodą CoMFA wymaga odpowiedniego przygotowania zbioru cząsteczek. Wszystkie cząsteczki poddaje się optymalizacji geometrii. Użyte konformacje nie mogą być przypadkowe. Jeżeli istnieje hipoteza farmakoforowa, dotycząca analizowanego zbioru, powinna być ona uwzględniona [14]. Konieczne jest również by analizowane związki wywoływały ten sam efekt biologiczny oraz by mechanizm wywoływania tego efektu był jednakowy. Do nakładania analizowanych cząsteczek wymagany jest wspólny motyw strukturalny. Odpowiednie nałożenie analizowanych cząsteczek ma kluczowe znaczenie w modelowaniu metodą CoMFA. Różnice między deskryptorami CoMFA obliczonymi dla dwóch różnych cząsteczek zależą nie tylko od różnic w ich strukturze lecz również od ich wzajemnej orientacji [15].

W przestrzeni zajmowanej przez analizowane cząsteczki konstruowana jest wirtualna trójwymiarowa siatka o określonym rozmiarze komórki. Wymiary siatki są tak dobierane, że obejmuje ona z odpowiednim zapasem wszystkie cząsteczki szeregu. Kolejno dla wszystkich analizowanych cząsteczek w węzłach siatki obliczane są wartości różnych pól

molekularnych. Zależnie od pola molekularnego w węzłach siatki umieszczane są odpowiednie sondy atomowe i obliczana jest energia oddziaływania między sondą a cząsteczką. Standardowo oblicza się pola oddziaływań elektrostatycznych i sterycznych.

Obliczone w węzłach siatki wartości pól molekularnych są numerycznym obrazem trójwymiarowych struktur chemicznych. Zależnie od ustalonego rozmiaru komórki oraz od liczby obliczonych pól każda cząsteczka jest opisywana przez kilka tysięcy zmiennych. Zestawienie obliczonych danych dla całego szeregu analizowanych cząsteczek skutkuje powstaniem macierzy (macierz $\boldsymbol{X}$), której wiersze odpowiadają kolejnym cząsteczkom (obiektom) a kolumny są zmiennymi opisującymi obiekty (deskryptory). Macierz $\boldsymbol{X}$ posiada zazwyczaj więcej zmiennych niż obiektów. Występują w niej również silne wzajemne korelacje zmiennych. Modelowanie zmiennej zależnej, y, za pomocą skorelowanych zmiennych niezależnych wymaga zastosowania odpowiednich metod statystycznych. W toku analizy CoMFA stosowana jest metoda PLS [16].

Wynikiem modelowania CoMFA jest ilościowa zależność między strukturą cząsteczek a aktywnością w postaci równia regresyjnego zawierającego tysiące współczynników korelacji. Znacznie bardziej użytecznym wynikiem analizy CoMFA są jednak wykresy konturowe. Przedstawiają one obszary przestrzeni zorientowane wokół analizowanych cząsteczek wskazujące pożądane lub niepożądane własności w danym obszarze (patrz rozdział 4.1, strona 41).

Największą słabością metody CoMFA jest duża zależność wyników od sposobu wzajemnego nałożenia cząsteczek [15]. Wirtualna siatka, w węzłach której obliczane są wartości pól molekularnych ma stałą orientację względem całego zbioru cząsteczek. Poszczególne węzły są przypisane odpowiednimi zmiennym. Nawet niewielka modyfikacja położenia pojedynczej cząsteczki (przesunięcie, obrót, również zmiana konformacji) może powodować drastyczną zmianę w wygenerowanym wektorze deskryptora. Węzły siatki znajdują się w całej przestrzeni zajmowanej przez cząsteczki, również w bezpośrednim sąsiedztwie atomów cząsteczek. Wartości pól obliczone w węzłach znajdujących się blisko atomów cząsteczek zmieniają się silnie nawet przy małej zmianie odległości węzła od atomu. Wynika to ze specyfiki stosowanych typów potencjałów. W zależności od pola są to potencjały typu Lennarda-Jonesa lub Coulomba. W pobliżu atomów ich wartości zmieniają się bardzo silnie, a gdy odległość zmierza do zera wartość tych potencjałów dąży do nieskończoności. Powoduje to, że obliczone

energie oddziaływań z sondami molekularnymi w niektórych węzłach siatki nie są wiarygodne. Trudności takie pokonuje się poprzez porównywanie obliczonych wartości z odpowiednią dla użytej sondy wartością progową. Wartości przekraczające próg są następnie odrzucane.

CoMFA jako pierwsza powszechnie stosowana metoda 3D-QSAR wyznacza pewien schemat tego typu analizy. Ustanawia też pewne minimalne standardy statystycznej jakości modeli. Wyniki modelowania innymi metodami są często porównywane z wynikami CoMFA. Wiele nowych metod jest ukierunkowanych na usunięcie wad metody CoMFA.

2.1.2 Metoda CoMSIA

Metoda CoMFA była modyfikowana i rozwijana przez wielu badaczy. Jedną z ciekawszych modyfikacji jest metoda CoMSIA, porównawcza analiza cząsteczkowych indeksów podobieństwa (ang. comparative molecular similarity indices analysis) [17]. Szereg analizowanych molekuł przygotowywany jest w analogiczny sposób jak w przypadku metody CoMFA. W węzłach siatki, dla każdej molekuły, obliczane są wartości różnych pól molekularnych. CoMSIA w odróżnieniu od CoMFA oblicza w węzłach siatki tzw. indeksy podobieństwa (ang. similarity indices):

$$A_{F,k}^{q}(j) = -\sum_{i=1}^{n} w_{\text{probe},k}\, w_{ik}\, \mathrm{e}^{-\alpha r_{iq}^{2}} \tag{2.2}$$

gdzie $A_{F,k}^{q}$ to wartość indeksu w węźle q, i oznacza atomy cząsteczki j, w_{ik} jest wartością wybranej własności k atomu i, $w_{\text{probe},k}$ jest wartością wybranej własności sondy molekularnej, q oznacza węzły siatki, r_{iq} jest odległością między atomem i a węzłem q, α jest współczynnikiem szerokości funkcji odległości.

Zastosowanie odpowiednich sond molekularnych pozwala obliczyć za pomocą wzoru (2.2) różne pola oddziaływań, np. pole oddziaływań sterycznych, elektrostatycznych, hydrofobowych oraz pola występowania donorów lub akceptorów wiązań wodorowych [18].

Zastąpienie potencjału Lennarda-Jonesa albo Coulomba (używanymi w metodzie CoMFA) funkcją odległości typu Gaussa powoduje, że pole obliczane w węzłach leżących bardzo blisko atomów cząsteczki, lub leżących na atomach cząsteczki nie przyjmuje

nieracjonalnie wielkich wartości co ma istotny wpływ na modelowanie [19]. W metodzie CoMSIA nie trzeba więc nakładać wartości progowych potencjału. Co więcej, zastosowana funkcja powoduje, że zmiana pola w pobliżu powierzchni cząsteczkowych nie jest tak gwałtowna jak w metodzie CoMFA przez co CoMSIA wykazuje większą niezależność od wzajemnego nałożenia analizowanych cząsteczek [18, 19]. Zwykle wyniki modelowania nie różnią się zasadniczo od wyników uzyskanych metodą CoMFA. [17].

2.1.3 Metoda SoMFA

Metoda porównawczej analizy samoorganizujących się pól molekularnych SoMFA (ang. self-organizing molecular field analysis) jest swoistym połączeniem różnych metod QSAR. Podobnie jak w metodzie CoMFA tworzona jest trójwymiarowa siatka obejmująca wszystkie analizowane związki. W węzłach siatki obliczane są wartości charakteryzujące kształt cząsteczek lub wartości potencjału elektrostatycznego. Kształt jest opisywany za pomocą dwóch wartości 0 i 1. Węzeł siatki przyjmuje wartość 1 jeśli znajduję się wewnątrz powierzchni van der Waalsa, w przeciwnym wypadku węzeł przyjmuje wartość 0 [20].

Najważniejszym krokiem analizy SoMFA jest przemnożenie wartości obliczonych w węzłach siatki dla poszczególnych cząsteczek przez ich wycentrowaną aktywność. Centrowanie polega na odjęciu od wszystkich aktywności wartości średniej.

Następnie trójwymiarowe siatki wygenerowane dla wszystkich cząsteczek są sumowane tworząc tzw. siatkę główną (ang. master grid). Wizualizacja siatki głównej daje obraz fragmentów przestrzeni o największym wkładzie w aktywność rozpatrywanych cząsteczek. Siatka główna służy również do prognozowania aktywności. W tym celu najpierw dla wszystkich cząsteczek obliczane są zbiorcze wartości właściwości używanych do tworzenia siatki głównej, tzw. zbiorcze wartości potencjału elektrostatycznego oraz kształtu cząsteczek. Wartości zbiorcze uzyskuje się przez sumowanie wartości wszystkich węzłów siatki przemnożonych przez wartości węzłów siatki głównej. Następnie metodą regresji liniowej w zbiorze wartości zbiorczych oraz wartości aktywności tworzony jest model QSAR.

Metoda SoMFA została z powodzeniem zastosowana do analizy zbioru steroidów wiążących globuliny CBG oraz szeregu sulfonamidowych inhibitorów endoteliny. Uzyskane modele charakteryzowały się wysoką zdolnością prognozowania [20].

2.1.4 Metoda CoMSA

Niedawno opisaną ciekawą metodą 3D-QSAR jest porównawcza analiza powierzchni cząsteczkowej CoMSA (ang. comparative molecular surface analysis) [21, 22, 23]. Łączy ona w sobie technikę samoorganizujących się map neuronowych SOM (ang. self-organizing maps – sieci neuronowe Kohonena [24, 25]) z analizą PLS [26].

Analiza QSAR (a także SAR) metodą CoMSA polega na porównywaniu powierzchni cząsteczkowych szeregu związków. W tym celu trójwymiarowe powierzchnie cząsteczkowe są przekształcane za pomocą sieci neuronowej w tzw. mapy porównawcze (ang. comparative maps). Mapy porównawcze są dwuwymiarowymi obrazami powierzchni cząsteczkowych. Ważną własnością map porównawczych jest zachowanie topologi obrazowanej powierzchni, dzięki czemu możliwa jest wizualizacja całej trójwymiarowej powierzchni za pomocą dwuwymiarowego obiektu.

Pierwszym etapem analizy CoMSA jest wytrenowanie sieci neuronowej przy użyciu powierzchni cząsteczki wzorcowej. Współrzędne punktów zebranych z powierzchni wzorca są wprowadzane do sieci neuronowej. Trening powoduje, że sieć samoczynnie uczy się rozpoznawania powierzchni cząsteczki wzorcowej.

Wytrenowana sieć jest w następnym etapie wykorzystywana do transformacji powierzchni analizowanego szeregu cząsteczek w szereg map porównawczych. Poszczególne neurony mapy są kolorowane średnią wartością potencjału obliczonego w punktach powierzchni, które trafiły do danego neuronu.

Uzyskane dwuwymiarowe obrazy powierzchni cząsteczkowych są następnie podawane analizie SAR oraz QSAR z wykorzystaniem metody PLS. Metoda CoMSA została z powodzeniem zastosowana do analizy QSAR wielu szeregów związków organicznych [21, 22, 23, 27, 28, 29, 30, 31, 32, 33]. Metoda CoMSA została zmodyfikowana przez Hasegawe. Modyfikacja polegała na zastosowaniu algorytmu PLS umożliwiającego analizę QSAR bez rozwijania map porównawczych na wektory (3-way PLS) [34].

2.1.5 Deskryptory WHIM

Zupełnie odmienne podejście kalkulacji deskryptorów charakteryzuje metody obliczające tzw. deskryptory WHIM czyli molekularne niezmienne holistyczne ważone deskryptory (ang. weighted holistic invariant molecular) [35, 36].

Obliczenie deskryptora WHIM można podzielić na kilka etapów. W pierwszym etapie centrowane są współrzędne atomów cząsteczki. Centrowanie wykonuje się względem ważonej średniej, każdemu atomowi *i* przypisuje się wagę w_i. Stosowane są cztery schematy wag: nieważony $w_i = 1$, mas atomowych $w_i = m_i$, objętości van der Waalsa $w_i = vdw_i$ oraz elektroujemności Mullikena $w_i = eln_i$. W następnym etapie wycentrowane współrzędne poddaje się analizie PCA uzyskując macierz wyników dla trzech czynników głównych. Na podstawie tej macierzy obliczane są następujące parametry statystyczne tworzące deskryptor WHIM [37]:

- Wariancje kolumn *m* – λ_m – czyli wartości własne PCA,
- Ułamki wartości własnych $\theta_m = \lambda_m / \sum_m \lambda_m$,
- Symetrie – $\gamma_m = \left| \sum_i (w_i t_{im}^3) / \sum_i w_i \right| * 1/\lambda_m^{3/2}$, gdzie t_{im} jest elementem macierzy wyników PCA a w_i jest wybraną wagą atomu *i*,
- Odwrotności krzywizn (ang. inverse function of the kurtosis) – $\eta_m = 1/\kappa_m$ gdzie $\kappa_m = [\sum_i (w_i t_{im}^4) / \sum_i w_i] * 1/\lambda_m^2$,
- Rozmiar całkowity (ang. total dimension) – $\tau = \lambda_1 + \lambda_2 + \lambda_3$,
- Czynnik acentryczny (ang. acentric factor) – $\omega = \theta_1 - \theta_3$,

Ponieważ tylko dwa pierwsze ułamki wartości własnych są wzajemnie niezależne, dla każdego schematu wag uzyskuje się więc 13 parametrów co daje w sumie 52 parametry.

Deskryptory WHIM pozwalają uzyskać modele, których jakość jest porównywalna z modelami CoMFA. Podstawową zaletą WHIM jest ich całkowita niezależność od sposobu nakładania cząsteczek oraz bardzo duża kompresja danych.

Tak zdefiniowane deskryptory można łatwo poddawać modyfikacji. Przykładem modyfikacji jest metoda G-WHIM (ang. grid-weighted holistic invariant molecular), w której współrzędne atomów są zastąpione współrzędnymi węzłów trójwymiarowej siatki [38]. W środku siatki umieszcza się cząsteczkę i za pomocą różnych sond w węzłach siatki oblicza się pole oddziaływań podobnie jak w metodzie CoMFA. Obliczone wartości są używane do centrowania współrzędnych węzłów. Inną modyfikacją tej metody jest MS-WHIM (ang. molecular surface WHIM) [39]. W tym przypadku do obliczania deskryptora

używane są punkty zebrane z powierzchni Connolyego, a różne własności powierzchniowe, takie jak potencjał elektrostatyczny, stanowią wagi tych punktów. Obie wspomniane metody charakteryzują się, podobnie jak oryginalna metoda WHIM, bardzo dużą kompresją danych a przy odpowiedniej gęstości próbkowania obliczone deskryptory są niezależne od wzajemnego nałożenia analizowanych cząsteczek [35].

2.1.6 Metoda AFMoC

Ciekawym zastosowaniem metodologii 3D-QSAR jest metoda AFMoC (ang. adaptation of fields for molecular comparision). Powstała ona przez połączenie zmodyfikowanej metody CoMFA oraz funkcji oceniającej (ang. scoring function) DrugScore [40].

Metoda DrugScore szacuje powinowactwo ligandów do białek na podstawie zbioru potencjałów typu atom-białko uzyskanych przez analizę oddziaływań występujących w różnych strukturach krystalograficznych. Używany przez nią zbiór oddziaływań typu atom-białko może być stosowany do oceny powinowactwa różnych ligandów do różnych białek.

Zastosowanie metodologii CoMFA do analizy grupy znanych ligandów umożliwia dopasowanie zbioru oddziaływań do wybranej proteiny.

Zaletą metody AFMoC jest możliwość stopniowego przekształcania ogólnego zbioru oddziaływań atom-białko do oddziaływań specyficznych dla wybranej proteiny w zależności od liczby dostępnych ligandów. Dzięki temu metoda AFMoC może być stosowana również do analizy małych zbiorów ligandów. Dodatkowo, dzięki zastąpieniu niespecyficznych oddziaływań oddziaływaniami atom-białko, interpretacja wyników modelowania AFMoC jest prostsza niż w przypadku metody CoMFA [40].

2.1.7 Metoda CoRSA

Metoda CoRSA, czyli porównawcza analiza powierzchni receptora (ang. comparative receptor surface analysis) sprowadza się do porównywania obrazów powierzchni szeregu cząsteczek [41].

Podobnie jak w przypadku innych metod 3D-QSAR przed przystąpieniem do analizy rozpatrywany zbiór cząsteczek należy poddać optymalizacji geometrii. Następnie z całego zbioru cząsteczek wybiera się kilka najbardziej aktywnych, zwykle od 1 do 5 molekuł.

Wybrane cząsteczki tworzą tzw. RGS (ang. receptor generation set). W metodzie CoRSA zakłada się, że na podstawie wybranego zbioru najbardziej aktywnych cząsteczek możliwe jest wygenerowanie wirtualnego receptora, zwanego również powierzchniowym modelem receptora (ang. receptor surface model – RSM) lub pseudo receptorem, którego obraz zbliżony jest do receptora rzeczywistego. Wygenerowany receptor reprezentowany jest, w odróżnieniu od receptora rzeczywistego nie przez atomy ale przez zbiór punktów zebranych z jego powierzchni. W każdym punkcie obliczane są różne własności np.: potencjał elektrostatyczny, ładunek cząstkowy, hydrofobowość a także zdolność tworzenia wiązań wodorowych. Istnieje wiele technik używanych do generowania RSM [42]. W przypadku metody CoRSA wykorzystywana jest metoda zaproponowana przez Hahn i Rogersa [43, 44, 45].

W następnym kroku dla każdego związku analizowanego szeregu oblicza się energię oddziaływania ze wszystkimi punktami wygenerowanego pseudo receptora [46]. W rezultacie cząsteczki reprezentowane są przez wektory opisujące różne oddziaływania z pseudo receptorem. Dodatkowo zazwyczaj przed obliczeniem energii oddziaływań geometrię umieszczonych w RSM cząsteczki poddaje się optymalizacji. W końcowym etapie do modelowania danych wykorzystuje się, podobnie jak w innych technikach 3D QSAR, metodę PLS.

2.2 Metody 4D-QSAR

Deskryptory stosowane w metodach 3D-QSAR są obliczane na podstawie trójwymiarowej struktury, która reprezentowana jest przez jedną konformację. W rzeczywistych procesach cząsteczki chemiczne ulegają stałym zmianom konformacyjnym a ich elastyczność może mieć kluczowe znaczenie dla aktywności.

Zmienność konformacyjna molekuł jest uwzględniania w modelowaniu 4D-QSAR. Metody tego typu obliczają deskryptory nie w oparciu o pojedynczą trójwymiarową strukturę ale na podstawie tzw. zbioru konformerów (ang. conformational ensemble profile – CEP). Dodatkowy czwarty wymiar oznacza więc zmienność konformacyjną trójwymiarowych struktur związków chemicznych. Wiele metod 4D-QSAR poza rożnymi konformerami uwzględnia również różne orientacje molekuł, formy tautomeryczne oraz ich różne uprotonowania [47].

2.2.1 Metody RI-4Q-QSAR

Opisana przez Hopfingera klasyczna analiza 4D-QSAR obejmuje następujące etapy: generowanie trójwymiarowych struktur, dynamika molekularna (generowanie zbioru konformerów) – generowanie CEP (ang. conformational ensemble profile), nakładanie, konstrukcja wirtualnej trójwymiarowej siatki obejmującej wszystkie CEP, definicja oddziaływań farmakoforowych (ang. interaction pharmacophore elements – IPEs), zliczanie atomów w komórkach siatki (ang. grid cell occupancy descriptors – GCODs), obliczanie modelu QSAR, poszukiwanie aktywnych konformacji [48].

W etapie generowania CEP w klasycznych metodach 4D-QSAR nie uwzględnia się wpływu receptora, z którym oddziałują badane związki. Z tego powodu metody te określane są jako niezależne od receptora (ang. receptor independent – RI-4D-QSAR).

Obliczony deskryptor molekularny jest zależny od rodzajów zastosowanych IPE (ang. interaction pharmacophore element) oraz od sposobu obliczania GCOD (ang. grid cell occupancy descriptor). Używane rodzaje IPE to: dowolny atom (ang. any type of atom), atom niepolarny (ang. nonpolar atom), atom polarny z ładunkiem dodatnim (ang. polar atoms of positive charge), atom polarny z ładunkiem ujemnym (ang. polar atoms of negative charge), akceptor wiązań wodorowych (ang. hydrogen bond acceptor), donor wiązań wodorowych (ang. hydrogen bond donor), atomy węgla i wodoru w układach aromatycznych (ang. aromatic carbons and hydrogens).

Generowanie deskryptora polegające na zliczaniu atomów w komórkach wirtualnej siatki – obliczanie GCOD – jest wykonywane na trzy różne sposoby. Obliczane są deskryptory typu absolutnego (ang. absolute), łącznego (ang. joint) oraz rozłącznego (ang. self). Deskryptor absolutny jest iloczynem logicznym występowania IPE w komórkach siatki. Do obliczenia deskryptora łącznego i rozłącznego wymagana jest cząsteczka referencyjna, względem której oblicza się deskryptory. Deskryptor typu łącznego jest iloczynem logicznym występowania IPE dla cząsteczki referencyjnej i dla cząsteczki, dla której obliczany jest deskryptor. Obliczanie deskryptora typu rozłącznego polega natomiast na pominięciu komórek siatki nie obsadzonych IPE w cząsteczce referencyjnej.

Ostatni etap analizy 4D-QSAR ma na celu znalezienie aktywnych konformacji badanych związków. Przez konformację aktywną rozumie się taką konformację jaką przybiera badany związek w rzeczywistym procesie oddziaływania z molekularnym makro celem. W tym celu dla wszystkich CEP wybierane są konformacje, których energia nie

wykracza poza ustalony próg względem średniej energii wszystkich konformacji CEP. Wybrane w ten sposób konformacje są następnie testowane względem uzyskanego modelu 4D-QSAR. Te, które prognozują najlepszą aktywność są uznawane za konformacje aktywne.

2.2.2 Metody RD-4Q-QSAR

Niedawno opisano w literaturze metody 4D-QSAR, które uwzględniają wpływ receptora podczas generowania CEP. Metody takie określa się mianem zależnych od receptora (ang. receptor dependent – RD-4D-QSAR) [49, 50, 51].

Wpływ receptora uwzględnia się przez symulacje dynamiki molekularnej po umieszczeniu ligandów w kieszeni receptora. Usytuowanie ligandów w kieszeni receptora określa się metodą dokowania molekularnego. Jeżeli struktura krystalograficzna receptora zawiera cząsteczkę ligandu w miejscu aktywnym dokowanie może być zrealizowane przez nakładanie analizowanych cząsteczek na znajdujący się w kieszeni ligand. W celu przyspieszenia obliczeń w czasie dynamiki molekularnej uwzględnia się zwykle jedynie miejsce aktywne wraz z najbliższym otoczeniem. Wygenerowane CEP są następnie analizowane podobnie jak w przypadku metod RI-4D-QSAR. Generowanie CEP w analizie RD-4D-QSAR umożliwia w pewnym stopniu symulację dopasowania indukowanego.

2.3 Metody 5D-QSAR

Dalszy rozwój metod QSAR związany jest z dokładniejszym uwzględnianiem wpływu receptora. Metody 5D-QSAR analogicznie do metod 4D-QSAR używają do reprezentacji ligandów CEP (ang. conformational ensemble profile) – przestrzeni konformacyjnej molekuł. Dodatkowy piąty wymiar odnosi się do uwzględniania różnych sposobów modelowania dopasowania indukowanego [52, 53].

2.3.1 Modelowanie dopasowania indukowanego

Program Quasar był pierwszym programem realizującym analizę 5D-QSAR (Quasar posiada również możliwość przeprowadzania modelowania 6D-QSAR – patrz również rozdział 2.4) [54, 55]. Obecnie dopasowanie indukowane symulowane jest na 6 rożnych sposobów: izotropowo (liniowe dopasowanie topologii), anizotropowo (dopasowanie w

oparciu o oddziaływania steryczne, elektrostatyczne, dopasowanie lipofilowe oraz dopasowanie w oparciu o analizę oddziaływań związanych z tworzeniem wiązań wodorowych) a także dopasowanie w oparciu o minimalizację energii [53, 56, 57].

Modelowanie QSAR w pakiecie Quasar opiera się na generowaniu quasi-atomowych modeli receptora [58]. W pierwszym kroku tworzony jest model receptora w postaci powierzchni van der Waalsa otaczającej wszystkie ligandy. Następnie powierzchnia modelu receptora odwzorowywana jest na tymczasowo tworzone powierzchnie poszczególnych ligandów. Odwzorowanie może być wykonane na 6 opisanych powyżej sposobów. Wartość średniego błędu kwadratowego (ang. root mean square – RMS) obliczona między zewnętrzną a wewnętrzną powierzchnią jest używana do szacowania energii dopasowania indukowanego.

W następnym kroku powierzchnie modelu receptora są losowo obsadzane różnymi własnościami atomowymi. Losowe obsadzenie jest następnie optymalizowane za pomocą algorytmów genetycznych. W wyniku tej procedury otrzymuje się zbiór modeli receptora. Dla wszystkich modeli obliczana jest energia wiązania, a na jej podstawie szacowana jest energia swobodna wiązania ligandów do receptora.

2.3.2 Dwupowłokowa reprezentacja receptora

Rozwinięciem modelowania dopasowania indukowanego jest metoda tzw. dwupowłokowej reprezentacji miejsca aktywnego (ang. dual-shell representation) [56]. Metoda dwupowłokowej reprezentacji miejsca aktywnego została pierwszy raz zaimplementowana w programie Raptor [59, 60, 61, 62]. Model receptora w tej metodzie stanowią dwie powłoki (powierzchnie). Powłoka wewnętrzna modeluje pola oddziaływań jakie są odczuwane przez cząsteczki dobrze dopasowane do miejsca aktywnego. Związki, które posiadają grupy sięgające w głąb receptora mogą jednak odczuwać inne oddziaływania będące wynikiem dopasowania indukowanego. Modelowanie tych oddziaływań jest realizowane przez drugą zewnętrzną powłokę.

Dopasowanie obu powłok do poszczególnych ligandów przebiega odmiennie niż w przypadku standardowej metody 5D-QSAR. Podczas dopasowywania uwzględniana jest zarówno topologia powłok jak i energie oddziaływań receptor-ligand. Uwzględniane są oddziaływania lipofilowe oraz związane z powstawaniem wiązań wodorowych. Następnie

za pomocą empirycznej funkcji (ang. empirical scoring function) szacowana jest energia swobodna wiązania ligand-receptor a na jej podstawie szacowane jest powinowactwo ligandów do receptora [56, 57].

2.4 Metody 6D-QSAR

Najnowszą metodą analizy QSAR jest metoda 6D-QSAR. Metoda ta została zaimplementowana w programie Quasar [47, 52, 54, 55, 58, 63]. Analiza 6D-QSAR jest najbardziej zaawansowaną metodą QSAR. Korzysta ona z trójwymiarowych struktur związków chemicznych (3D-QSAR) do generowania CEP – zbioru konformerów, tautomerów a także różnych stanów protonacji (4D-QSAR). Wygenerowane CEP są używane do tworzenia quasi-atomowych modeli receptora, które są następnie dopasowywane do wszystkich struktur na 6 różnych sposobów (5D-QSAR). Wprowadzenie szóstego wymiaru (6D-QSAR) wiąże się z możliwością uwzględniania kilku różnych modeli solwatacji. Równanie używane do obliczania energii wiązania ligand – receptor w metodzie 6D-QSAR zostało więc uzupełnione o dodatkowy wyraz odpowiadający zmianie energii solwatacji [64].

Symulacja różnych modeli solwatacji może być wykonana w sposób bezpośredni lub pośredni. W pierwszym przypadku powierzchnie modeli receptora są uzupełniane o dodatkowe właściwości związane z efektami solwatacji. Rozmieszczenie tych obszarów jest, podobnie jak w przypadku analizy 5D-QSAR, optymalizowane przy użyciu algorytmów genetycznych. Pośrednia symulacja solwatacji polega natomiast na niezależnym skalowaniu dla każdego modelu receptora wyrazów odpowiadających zmianie energii solwatacji. Skalowanie odzwierciedla różnice w dostępności poszczególnych modeli receptora dla cząsteczek rozpuszczalnika. Wartości wag przypisywane poszczególnym modelom są również optymalizowane przy użyciu algorytmów genetycznych. Według autorów metody podejście pośrednie pozwala uzyskać lepsze wyniki [64].

3 Problemy przetwarzania danych w modelowaniu m-QSAR

Modelowanie wielowymiarowych zależności QSAR (m-QSAR), sprowadza się do analizy wielowymiarowej macierzy zmiennych opisujących struktury analizowanych cząsteczek w sposób liczbowy. Każda liczbowa reprezentacja struktur cząsteczkowych stosowana w metodach QSAR nazywana jest deskryptorem. Dane QSAR najczęściej posiadają układ, w którym wiersze macierzy reprezentują poszczególne cząsteczki a kolumny przechowują wartości deskryptorów. Istnieje wiele rodzajów deskryptorów stosowanych w modelowaniu QSAR. Podstawowym typem są deskryptory topologiczne obliczane na podstawie dwuwymiarowego grafu cząsteczki. Deskryptory stosowane w wielowymiarowych metodach QSAR są raczej zbiorem wielu zmiennych niż pojedynczą kolumną. Przykładowo, zastosowanie siatki o gęstości 1 Å i wymiarach 10x10x10 Å w metodzie CoMFA generuje macierz zawierającą deskryptor liczący 1000 kolumn. W toku dalszej analizy część kolumn jest zazwyczaj eliminowana, pozostałe natomiast tworzą wielowymiarowy deskryptor QSAR. Podobną procedurę prowadzi się w przypadku innych metod modelowania m-QSAR (patrz również rozdział 2, strona 8).

Podstawą szeroko rozumianego modelowania QSAR jest odwzorowanie macierzy deskryptorów, macierzy $\boldsymbol{X}$, na macierz $\boldsymbol{Y}$ (lub wektor $\boldsymbol{y}$) opisującą aktywności cząsteczek. Organizacja macierzy aktywności jest analogiczna do organizacji macierzy deskryptorów. Wiersze macierzy $\boldsymbol{Y}$ odpowiadają cząsteczkom, kolumny natomiast zawierają aktywności. W praktyce najczęściej stosowany jest tylko jeden rodzaj aktywności – wówczas macierz $\boldsymbol{Y}$ staje się wektorem kolumnowym $\boldsymbol{y}$.

Ilościowe powiązanie deskryptorów z aktywnością wymaga zastosowania odpowiednich metod statystycznych. Wielokrotna regresja liniowa MLR (ang. multiple linear regression) ze względu na specyfikę danych QSAR nie znajduje praktycznego zastosowania w modelowaniu m-QSAR gdyż w przypadku tych danych konieczne jest użycie takich metod jak regresja głównych składowych PCR (ang. principal components regression) lub cząstkowa regresja najmniejszych kwadratów PLS (ang. partial least squares). W toku analizy szerokie zastosowanie znajdują również metody eliminacji i wyboru zmiennych na przykład algorytmy genetyczne GA (ang. genetic algorithm) [65, 66, 67, 68, 69, 70, 71, 72]. Opisano ostatnio zastosowanie do tego celu metody UVE [31, 73, 74].

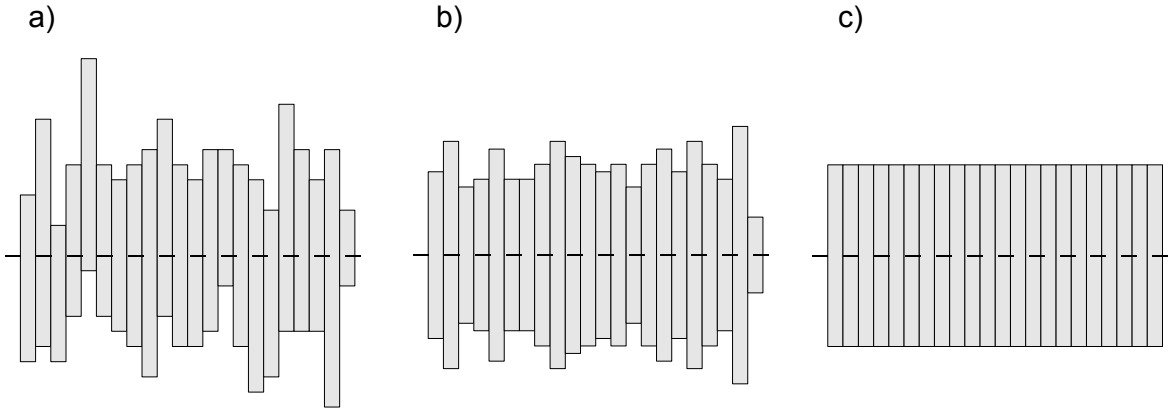
3.1 Wstępne przetwarzanie danych

Przed modelowaniem dane QSAR należy odpowiednio przygotować. Zmienne tworzące macierz $\boldsymbol{X}$ mogą się znacznie różnić zakresem wartości. Zdecydowana większość stosowanych metody analizy danych wymaga unormowania zakresów zmiennych. W praktyce QSAR stosuje się dwa sposoby – centrowanie oraz standaryzację (autoskalowanie). Rysunek 3.1 przedstawia schemat działania centrowania i standaryzacji danych.

Centrowanie danych powoduje przesunięcie zakresów zmiennych w taki sposób, że wartości średnie wszystkich zmiennych są równe zero. W tym celu od każdej zmiennej odejmuje się jej wartość średnią:

$$\boldsymbol{X}^C = \boldsymbol{X} - \bar{\boldsymbol{x}} \cdot \boldsymbol{I} \tag{3.1}$$

gdzie $\boldsymbol{X}^C$ jest wycentrowaną macierzą $\boldsymbol{X}$, $\bar{\boldsymbol{x}}$ jest wektorem wartości średnich wszystkich kolumn macierzy $\boldsymbol{X}$ a $\boldsymbol{I}$ jest macierzą jednostkową o wymiarach macierzy $\boldsymbol{X}$.



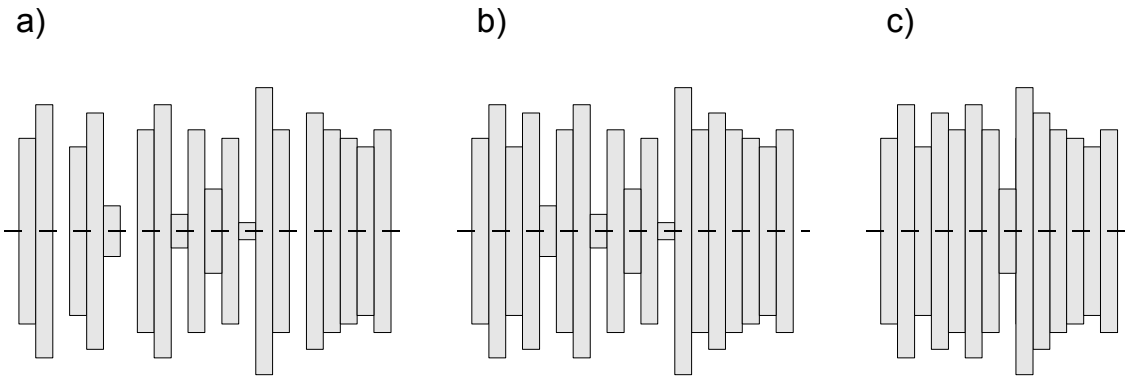
Rysunek 3.1 Centrowanie oraz standardyzacja danych. Zmienne są przedstawione w postaci pionowych pasków. Wielkość pasków wyraża wariancję danych a wzajemne położenie odpowiada zakresowi wartości. Część a przedstawia oryginalne dane, część b dane po centrowaniu, część c dane po standardyzacji.

W wielu przypadkach zmienne różniące się znacznie wartością wariancji powinny posiadać tę samą wagę. Standaryzacja (autoskalowanie) powoduje zrównanie wariancji wszystkich zmiennych. W pierwszej kolejności zmienne są centrowane a następnie dzielone przez wartość odchylenia standardowego:

$$\boldsymbol{X}^S = \frac{\boldsymbol{X}^C}{\mathrm{std}(\boldsymbol{X}) \cdot \boldsymbol{I}} \tag{3.2}$$

gdzie $\boldsymbol{X}^S$ jest wystandaryzowaną macierzą $\boldsymbol{X}$, funkcja std( ) zwraca wektor wartości średnich poszczególnych kolumn macierzy $\boldsymbol{X}$ a $\boldsymbol{I}$ jest macierzą jednostkową o wymiarach macierzy $\boldsymbol{X}$. Nie zawsze standaryzacja danych przynosi korzystne wyniki. Jeżeli w analizowanej macierzy są zmienne niosące wyraźny sygnał oraz zmienne zawierające jedynie szum tła (np. w przypadku analizy widm molekularnych) standaryzacja zrówna wariancję rzeczywistych sygnałów z szumem. Informacja zostaje wówczas przysłonięta przez szumy [75].

Kolejnym krokiem wstępnego przygotowania danych jest usunięcie z macierzy kolumn posiadających zerową lub znikomą wariancję. Kolumny o zerowej wariancji w analizie QSAR są zwykle kolumnami zawierającymi wyłącznie zera. Usunięcie kolumn o zerowej wariancji nie zmienia wyników modelowania pozwala zaś w wielu przypadkach znacznie przyśpieszyć obliczenia. Czasami usuwa się także kolumny o małej ale niezerowej wariancji. Jest to jednak ryzykowne ponieważ takie zmienne niosą pewną informację. Rysunek 3.2 przedstawia schematycznie usuwanie kolumn o zerowej oraz o bardzo małej wariancji.



Rysunek 3.2 Usuwanie pustych kolumn. Zmienne są przedstawione w postaci pionowych pasków ich wielkość wyraża wariancję danych. Część a przedstawia oryginalne dane, część b dane po usunięciu zerowych kolumn, część c dane po usunięciu kolumn o niskiej wariancji.

3.2 Wielokrotna regresja liniowa – MLR

Modelowanie aktywności w metodach QSAR wymaga zwykle zastosowania regresji wielokrotnej (regresji wielorakiej). Jedynie w przypadku zastosowania do opisu struktur pojedynczego deskryptora możliwe jest zastosowanie prostej regresji dwóch zmiennych. Ogólny wzór łączący deskryptor z aktywnością można wyrazić następująco:

$$y = b_0 + x \cdot b \tag{3.3}$$

gdzie y jest modelowanym efektem, x wyraża w sposób liczbowy strukturę związku, b jest współczynnikiem regresji a b_0 oznacza współczynnik regresji dla wyrazu wolnego. Natomiast w przypadku wielowymiarowych danych wzór (3.3) należy przedstawić w następującej postaci:

$$y_n = b_0 + \sum_m x_{n\,m} \cdot b_m \tag{3.4}$$

gdzie y_n jest elementem n wektora $\boldsymbol{y}$ zawierającego aktywności, b_m jest elementem m wektora $\boldsymbol{b}$ zawierającego współczynniki regresji, $x_{n\,m}$ oznacza element macierzy $\boldsymbol{X}$. Jest to podstawowe równanie regresyjne stosowane w analizie 3D-QSAR. Obliczenie współczynników $\boldsymbol{b}$ pozwala ilościowo powiązać deskryptor z aktywnością. W wypadku modelowania rzeczywistych efektów wzór (3.4) powinien być jeszcze uzupełniony o błąd obliczanej wartości e_n.

Istnieje wiele metod obliczania współczynników $\boldsymbol{b}$. Podstawową jest metoda wielokrotnej regresji liniowej MLR (ang. multiple linear regression) [76]. Analogicznie do regresji jednowymiarowej równanie regresyjne może być zapisane w postaci macierzowej:

$$\boldsymbol{y} = \boldsymbol{X} \cdot \boldsymbol{b} \tag{3.5}$$

gdzie $\boldsymbol{y}$ jest wektorem zmiennych zależnych, $\boldsymbol{X}$ jest macierzą zmiennych niezależnych opisujących obiekty a $\boldsymbol{b}$ jest wektorem współczynników regresji. Dołączenie do macierzy $\boldsymbol{X}$ dodatkowej kolumny jedynek pozwala również zawrzeć we wzorze (3.5) współczynnik b_0. Obliczenie $\boldsymbol{b}$ możliwe jest za pomocą następującego przekształcenia:

$$\boldsymbol{b} = (\boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}' \cdot \boldsymbol{y} \tag{3.6}$$

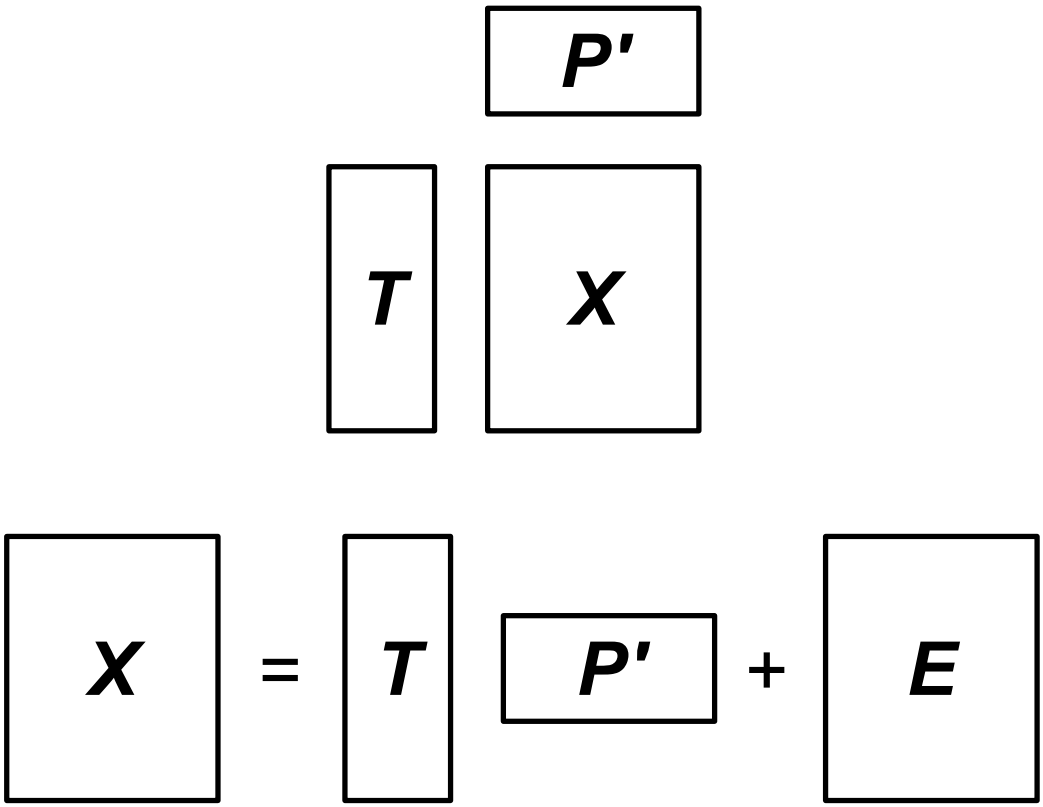
Metoda MLR posiada istotne ograniczenie w modelowaniu QSAR. Jeżeli w macierzy $\boldsymbol{X}$ występują silne korelacje między zmiennymi (kolumnami) nie jest możliwe poprawne odwrócenie macierzy ($\boldsymbol{X}' \cdot \boldsymbol{X}$). Wzór (3.6) nie może wówczas być użyty do

obliczenia współczynników ***b***. Występowanie takich korelacji w przypadku wielowymiarowych metod QSAR jest bardzo powszechne, dlatego konieczne jest stosowanie innych metod, np. PCR (ang. pricipal component regression) lub PLS (ang. partial least squares) – zobacz rozdziały 3.3.2 oraz 3.4, strony 29 oraz 30.

3.3 Analiza czynników głównych – PCA

W metodzie analizy czynników głównych (ang. principal component analysis – PCA) oblicza się na podstawie macierzy danych tzw. czynniki główne [77, 78]. Czynniki główne są kombinacją liniową pierwotnych zmiennych, utworzoną w taki sposób by maksymalizować opis wariancji danych. Pierwotna macierz jest dekomponowana na dwie, tj. macierz wyników oraz macierz wag. Jeżeli dekompozycja macierzy jest całkowita, macierze wag i wyników mają maksymalną możliwą liczbę czynników, wówczas iloczyn tych dwóch macierzy odtwarza pierwotną macierz danych. W wypadku kiedy w modelach wykorzystuje się tylko kilka pierwszych czynników pierwotna macierz jest odtwarzana z pewnym błędem. Rysunek 3.3 ilustruje schematycznie zależność między macierzą danych a macierzami wyników, wag i błędów.

Macierz ***T*** jest takim odpowiednikiem macierzy ***X***, że jej kolumny są ortogonalne. Każda kolumna tej macierzy jest czynnikiem głównym utworzonym w ten sposób by maksymalnie obrazować wariancję macierzy ***X***. Kolejne czynniki główne opisują pewien ułamek całkowitej wariancji, przy czym wraz z każdym kolejnym czynnikiem ułamek ten maleje. Maksymalna liczba czynników zależy od rzędu macierzy ***X***. Użycie maksymalnej liczby czynników, równej rzędowi macierzy ***X***, gwarantuje, że macierz ***T*** w pełni opisuje wariancję danych macierzy ***X***. W praktycznej analizie PCA używanych jest tylko kilka lub kilkanaście pierwszych czynników. Uznaje się, że wariancja opisywana przez pozostałe czynniki jest szumem informacyjnym lub nie jest konieczna do modelowania. Natomiast wiersze macierzy ***T*** odpowiadają wierszom macierzy ***X***, czyli obiektom. Wszelkie relacje występujące między obiektami w pierwotnej macierzy takie jak wzajemna odległość są również zachowane w macierzy ***T***.

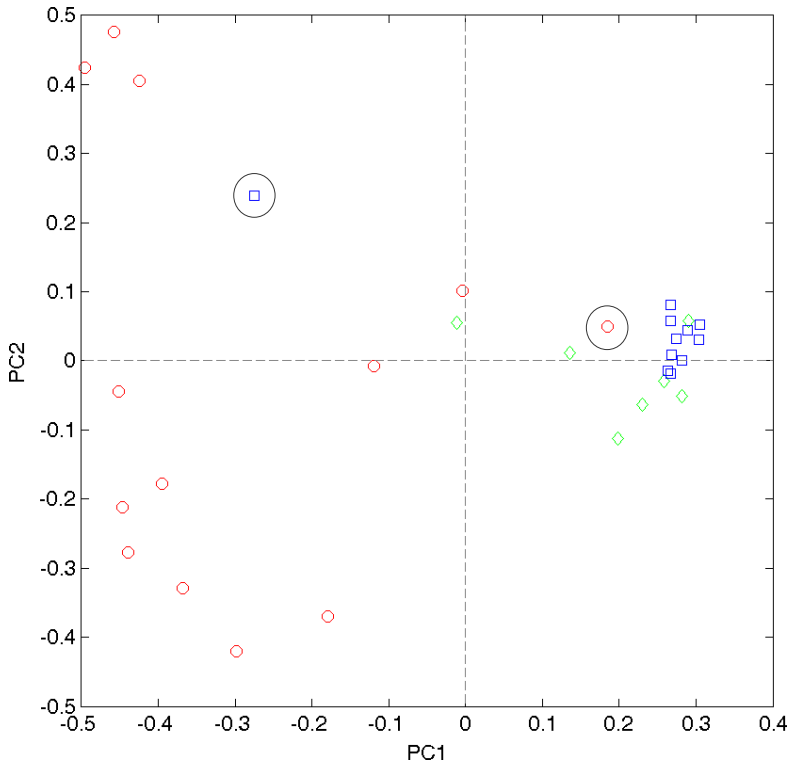


Rysunek 3.3 Schemat blokowy dekompozycji macierzy danych ***X*** na macierz wyników ***T*** oraz macierz wag ***P***. Iloczyn macierzy wag i wyników odtwarza macierz ***X*** z ewentualnym uwzględnieniem macierzy błędów ***E***.

Oryginalne zmienne są charakteryzowane w analizie PCA przez kolumny macierzy ***P***. Każda kolumna tej macierzy podobnie jak w przypadku macierzy ***T*** nosi nazwę czynnika głównego. Liczba kolumn macierzy ***P*** jest równa liczbie kolumn macierzy ***T***. Analogicznie jak w przypadku macierzy ***T*** wszelkie relacje występujące między oryginalnymi zmiennymi są zachowane w macierzy ***P***. Dzięki temu w prosty sposób można wizualizować korelacje występujące między zmiennymi.

3.3.1 Wizualizacja czynników głównych

Czynniki główne uzyskane przez dekompozycję macierzy można wykorzystać w celu wizualizacji zależności występujących między obiektami oraz zmiennymi. Rysunek 3.4 przedstawia projekcję obiektów na płaszczyznę zdefiniowaną czynnikami PC1 i PC2 uzyskaną w wyniku prostej analizy danych QSAR uzyskanych metodą s-CoMSA grupy pochodnych steroidowych o aktywności CBG (patrz rozdział 5.4.1.1, strona 52). Każdy punkt wykresu został opatrzony etykietą informującą o powinowactwie do globuliny wiążącej kortyzol – patrz opis pod rysunkiem. Pomimo tego, że czynniki główne PC1 oraz PC2 nie uwzględniają wartości powinowactwa, rozkład punktów w przestrzeni PC1/PC2 ilustruje zmiany powinowactwa CBG, które wzrasta ze wzrostem PC1. Oznacza to, że różnice strukturalne w budowie cząsteczek tłumaczą także ich wyraźne powinowactwo wobec CBG. Dwie pochodne zaznaczone obwódką (jedna po lewej stronie wykresu, druga po prawej, odpowiednio związki **s22** oraz **s31**) nie poddają się wyraźnie takiemu opisowi.



Rysunek 3.4 Przykładowa projekcja obiektów na płaszczyznę zdefiniowaną czynnikami PC1 i PC2 (ang score plot). Analiza PCA została przeprowadzona dla grupy pochodnych steroidowych o aktywności CBG (patrz rozdział 5.4.1.1, strona 52). Związki o niskim, średnim oraz wysokim powinowactwie oznaczono odpowiednio czerwonymi kółkami, zielonymi rombami oraz niebieskimi kwadratami. Obwódką zaznaczono pochodne **s22** i **s31** różniące się od pozostałych związków szeregu.

3.3.2 Regresja czynników głównych – metoda PCR

Metoda PCR (ang. principal component regression) jest modyfikacją metody MLR polegającą na zastosowaniu macierzy czynników głównych $\boldsymbol{T}$ w miejsce macierzy $\boldsymbol{X}$ [76]. Obliczenie współczynników $\boldsymbol{b}$ jest możliwe analogicznie jak we wzorze (3.6):

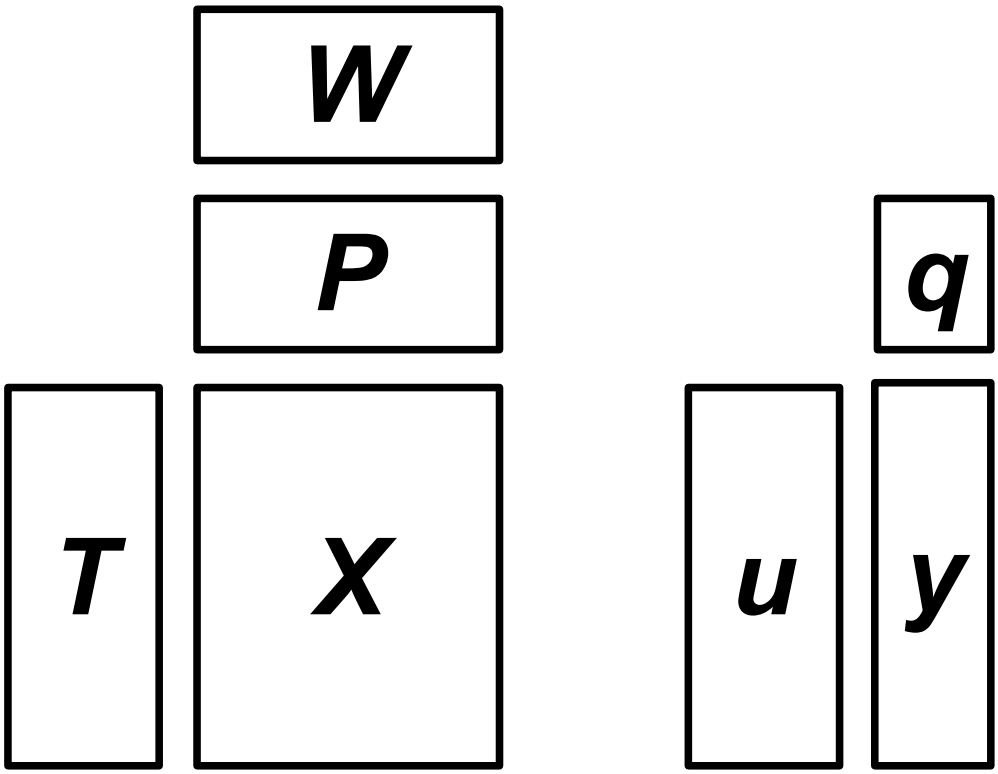
$$\boldsymbol{b}=(\boldsymbol{T}'\cdot\boldsymbol{T})^{-1}\cdot\boldsymbol{T}'\cdot\boldsymbol{y} \tag{3.7}$$

Odwrócenie macierzy ($\boldsymbol{T}'\cdot\boldsymbol{T}$) jest możliwe ponieważ czynniki główne są wzajemnie ortogonalne.

Zastosowanie czynników głównych powoduje jednak, że zmienne o bardzo dużej wariancji mają bardzo silny wpływ na kształt modelu. W analizie QSAR takie zmienne nie zawsze są skorelowane z modelowaną aktywnością. Efekt ten jest widoczny zwłaszcza w przypadku tzw. deskryptorów charakterystycznych (ang. fingerprint descriptors). Dlatego przed modelowaniem metodą PCR często konieczna jest preselekcja zmiennych. Wadą metody PCR jest również konieczność użycia względnie dużych zbiorów obiektów w celu otrzymania wiarygodnych modeli [79].

3.4 Regresja częściowych najmniejszych kwadratów – PLS

Regresja metodą najmniejszych częściowych kwadratów – PLS (ang. partial least squares) opiera się podobnie jak w przypadku metody PCR na dekompozycji macierzy ***X*** [16, 80]. Tworzone są tzw. ukryte czynniki (ang. latent components) będące kombinacją liniową oryginalnych zmiennych. Nowe zmienne, inaczej niż w przypadku czynników głównych PCA, maksymalizują kowariancję między macierzą ***X*** a wektorem ***y***. Na ich podstawie w iteracyjnej procedurze obliczane są współczynniki korelacji [81, 82, 83]. Rysunek 3.5 ilustruje schematycznie dekompozycję macierzy w metodzie PLS [83].



Rysunek 3.5 Schemat blokowy dekompozycji PLS. Macierz danych ***X*** jest dekomponowana na macierz wyników ***T***, macierz ładunków czynnikowych (obciążeń) ***P*** oraz macierz wag ***W***. Wektor ***y*** jest dekomponowany na wektory wyników ***u*** i ładunków czynnikowych (obciążeń) ***q***. Iloczyn odpowiednich macierzy i wektorów wyników i obciążeń daje pierwotną macierz lub wektor.

Metoda PLS jest rutynową metodą modelowania wielowymiarowych danych QSAR. Jej zaletą jest znacząca kompresja danych. Zastosowanie metody PLS w modelowaniu QSAR umożliwiło *de facto* rozwój tych metod. Jest to obecnie najszerzej stosowana metoda regresyjna w modelowaniu m-QSAR.

3.4.1 Walidacja modelu

Specyfika danych QSAR sprawia, że walidacja statystycznej reprezentatywności nie może być dokonana jedynie przez szacowanie dopasowania za pomocą średniego błędu kwadratowego *RMS* (ang. root mean sqauare) wzór (3.8) lub kwadratu współczynnika korelacji liniowej r^2 wzór (3.9):

$$RMS = \sqrt{\frac{\sum_{i=1}^{n} (y_i^p - y_i)^2}{n}} \tag{3.8}$$

$$r^2 = 1 - \frac{\sum_{i=1}^{n} (y_i^p - y_i)^2}{\sum_{i=1}^{n} (\bar{y} - y_i)^2} \tag{3.9}$$

gdzie y_i^p jest elementem i wektora $\boldsymbol{y}^p$ zawierającego prognozowane wartości zmiennej zależnej, y_i jest elementem i wektora $\boldsymbol{y}$ zawierającego oryginalne wartości zmiennej zależnej, n jest liczbą obiektów, $\bar{y}$ jest średnią wartością zmiennej zależnej. Obydwie wartości w prosty sposób szacują wzajemne dopasowanie danych i modelu. Zaletą parametru r^2 jest brak wymiaru (o ile $\boldsymbol{y}$ posiada wymiar) i co za tym idzie niezależność od zakresu wartości zmiennej zależnej.

W analizie m-QSAR walidację modeli przeprowadza się najczęściej przez obliczanie parametru q^2. Jest to walidowana krzyżowo wersja parametru r^2 – wzór (3.10). Walidacja krzyżowa (walidacja naprzemienna) zwana jest potocznie z języka angielskiego cross-walidacją (ang. cross-validation). W literaturze często stosuje się symbol q^2_{CV}.

Zbiór obiektów służący do uzyskania modelu jest w czasie walidacji krzyżowej wielokrotnie dzielony na dwa podzbiory. Jeden podzbiór jest używany do wygenerowania tymczasowego modelu (w normalnym toku obranej metody modelowania). Model ten jest następnie testowany na obiektach tworzących drugi podzbiór.

Przewidziane w czasie testowania wartości zmiennej zależnej służą następnie do obliczania wartości q^2_{CV}:

$$q^2_{CV} = 1 - \frac{\sum_{i=1}^{n} (y_i^{CV} - y_i)^2}{\sum_{i=1}^{n} (\bar{y} - y_i)^2} \tag{3.10}$$

gdzie y_i^{CV} jest elementem i wektora $\boldsymbol{y}^{CV}$ zawierającego wartości zmiennej zależnej oszacowane w czasie walidacji krzyżowej.

Innym często stosowanym parametrem jest walidowany krzyżowo błąd standardowy s_{CV}:

$$s_{CV} = \sqrt{\frac{\sum_{i=1}^{n} (y_i^{CV} - y_i)^2}{n - a^{opt} - 1}} \tag{3.11}$$

gdzie a^{opt} oznacza optymalną liczbę ukrytych czynników PLS.

3.4.1.1 Walidacja krzyżowa LOO

Jeżeli podczas walidacji krzyżowej proporcje podziału zbioru obiektów wynoszą n-1÷1 jest to walidacja typu LOO (ang. leave one out). Jest to najczęściej stosowany typ walidacji krzyżowej w analizie QSAR. Walidacja LOO jest czasochłonna, jej zaletą jest jednak duża wiarygodność wyników.

3.4.1.2 Walidacja krzyżowa LSO

Jeżeli proporcje podziału zbioru obiektów podczas walidacji krzyżowej wynoszą $(n\text{-}n_{CV})\div n$ gdzie n_{CV} jest liczbą całkowitą z zakresu (2, n) jest to walidacja typu LSO (ang. leave several out). Walidacja tego typu jest szybsza niż walidacja LOO. Uzyskane wyniki są jednak mniej wiarygodne.

3.4.2 Kompleksowość modelu

Istotnym problemem w analizie PLS jest określenie kompleksowości modelu czyli ustalenie liczby ukrytych czynników umożliwiającej uzyskanie optymalnego modelu. Czynniki PLS są konstruowane tak by maksymalizować kowariancję między macierzą $\boldsymbol{X}$ a wektorem $\boldsymbol{y}$. Większa liczba czynników powoduje więc, że model uwzględnia większy procent kowariancji. Podobnie jednak jak w przypadku metody PCA cześć kowariancji pochodzi zwykle od szumu. Model uwzględniający szum charakteryzuje się tzw. niską

zdolnością prognozowania. Oznacza to, że chociaż dobrze dopasowuje się do danych macierzy $\boldsymbol{X}$ nie jest zdolny do prawidłowego obliczania odpowiedzi $\boldsymbol{y}$ dla danych zewnętrznych (nie uwzględnionych na etapie modelowania). Modele takie określa się jako przeuczone.

Ustalenie odpowiedniej kompleksowości wykonuje się przez optymalizację parametru *PRESS* (ang. predictive residual sum of squares) obliczanego dla kolejnych kompleksowości:

$$PRESS(a)=\sum_{i=1}^{n}(y_i^a-y_i)^2$$
$$PRESS=\left[PRESS(1),\quad \ldots,\quad PRESS(a),\quad \ldots,\quad PRESS(A)\right] \tag{3.12}$$

gdzie a oznacza kompleksowość, A jest maksymalną użytą kompleksowością, $\boldsymbol{y}^a$ jest wektorem zmiennych zależnych oszacowanych dla kompleksowości a, $\boldsymbol{y}$ jest natomiast wektorem oryginalnych zmiennych zależnych. W iteracyjnym algorytmie PLS w każdym kroku zwiększana jest kompleksowość modelu od kompleksowości 1 do kompleksowości maksymalnej, równej zwykle liczbie z zakresu od 5 do 20. Wartość *PRESS* jest błędem oszacowania wektora $\boldsymbol{y}$. Na jego podstawie obliczany jest parametr *RMSCV* (ang. root mean square error of cross-validation) służący do określenia prawidłowej kompleksowości modelu. *RMSCV* oblicza się z następującego wzoru:

$$RMSCV(a)=\frac{PRESS(a)}{n}$$
$$RMSCV=\left[RMSCV(1),\quad \ldots,\quad RMSCV(a),\quad \ldots,\quad RMSCV(A)\right] \tag{3.13}$$

gdzie n jest liczbą obiektów. W celu zwiększenia nacisku na wybór niższych kompleksowości wzór (3.13) można zmodyfikować uwzględniając liczbę użytych czynników:

$$RMSCV(a)=\frac{PRESS(a)}{n-a-1}$$
$$RMSCV=\left[RMSCV(1),\quad \ldots,\quad RMSCV(a),\quad \ldots,\quad RMSCV(A)\right] \tag{3.14}$$

Na podstawie *RMSCV* określana jest kompleksowość modelu. Za optymalną kompleksowość uważa się taką, dla której wartość *RMSCV* jest najmniejsza. Alternatywną metodą określania optymalnej kompleksowości jest poszukiwanie pierwszego minimum parametru *RMSCV*. Symbolem używanym w tej pracy do określenia maksymalnej użytej liczby czynników jest *A*, znaleziona optymalna liczba czynników jest z kolei oznaczana symbolem a^{opt}.

3.4.3 Stabilność zmiennych

Miarą stabilności zmiennych jest niezależność współczynników regresji od zestawu obiektów użytych do konstrukcji modelu. Jeżeli drobna zmiana zestawu obiektów służących do uzyskania modelu silnie wpływa na wartość współczynnika regresji współczynnik ten jest niestabilny. Jeżeli natomiast wartość współczynnika nie ulega silnej zmianie jest on współczynnikiem stabilnym. Zmienne, którym odpowiadają stabilne współczynniki regresji są również nazywane stabilnymi. Te, którym odpowiadają niestabilne współczynniki są nazywane zmiennymi niestabilnymi [31].

Stabilność zmiennych może być ustalona w czasie walidacji krzyżowej. Wartości współczynników $\boldsymbol{b}$ wszystkich tymczasowych modeli są zapamiętywane w macierzy $\boldsymbol{BB}$. Zaproponowano kilka metod obliczania stabilności zmiennych. Najprostszy sposób polega na podzieleniu średniej wartości współczynników $\boldsymbol{b}$ przez odpowiadające im odchylenia standardowe – wzór (3.15). Zastąpienie średniej medianą a odchylenia standardowego interkwartylem ogranicza wpływ destabilizacyjny obiektów odległych. Jest to tzw. elastyczna stabilność (ang. robust) – wzór (3.16).

$$\boldsymbol{sv} = \mathrm{mean}(\boldsymbol{BB}) ./ \mathrm{std}(\boldsymbol{BB}) \tag{3.15}$$

$$\boldsymbol{sv} = \mathrm{median}(\boldsymbol{BB}) ./ \mathrm{iqr}(\boldsymbol{BB}) \tag{3.16}$$

Można również w miejsce średniej użyć ostatecznie uzyskaną wartości współczynników $\boldsymbol{b}$:

$$\boldsymbol{sv} = \boldsymbol{b} ./ \mathrm{std}(\boldsymbol{BB}) \tag{3.17}$$

We wzorach (3.15, 3.16, 3.17) $\boldsymbol{sv}$ oznacza wektor stabilności, symbol ./ oznacza dzielenie Hadamarda (element przez element), funkcje mean( ), std( ), median( ), iqr( ) są kolejno funkcjami zwracającymi wartość średnią, odchylenie standardowe, medianę oraz interkwartyl (odległość między 25 a 75 percentylem).

3.5 Walidacja modeli dla zbioru testowego

Parametry testujące zdolność prognozowania modeli takie jak q^2_{CV}, s_{CV}, r^2, *RMS* określają dopasowanie modelu do danych treningowych. Oznacza to, że wiarygodność modeli jest testowana na obiektach używanych do konstrukcji modelu. Parametry q^2_{CV} oraz s_{CV} są bardziej wiarygodne z uwagi na zastosowanie do ich obliczenia walidacji krzyżowej. Dla dużych zbiorów możliwe jest wyodrębnienie spośród obiektów zewnętrznego zbioru testowego nie używanego do konstrukcji modelu. Testowanie zdolności prognozowania na podstawie zbioru zewnętrznego jest bardziej wiarygodne. Stosowane w tym celu parametry są analogiczne do parametrów r^2 i *RMS*. Najczęściej stosowane to standardowy błąd prognozowania *SDEP* (ang. standard deviation error of prediction) oraz kwadrat współczynnika dopasowania zbioru testowego r_t^2 (r^2 *test*):

$$SDEP = \sqrt{\frac{\sum_{i=1}^{n_t} (y_i^{pt} - y_i^t)^2}{n_t}} \tag{3.18}$$

$$r_t^2 = 1 - \frac{\sum_{i=1}^{n_t} (y_i^{pt} - y_i^t)^2}{\sum_{i=1}^{n_t} (\bar{y}^t - y_i^t)^2} \tag{3.19}$$

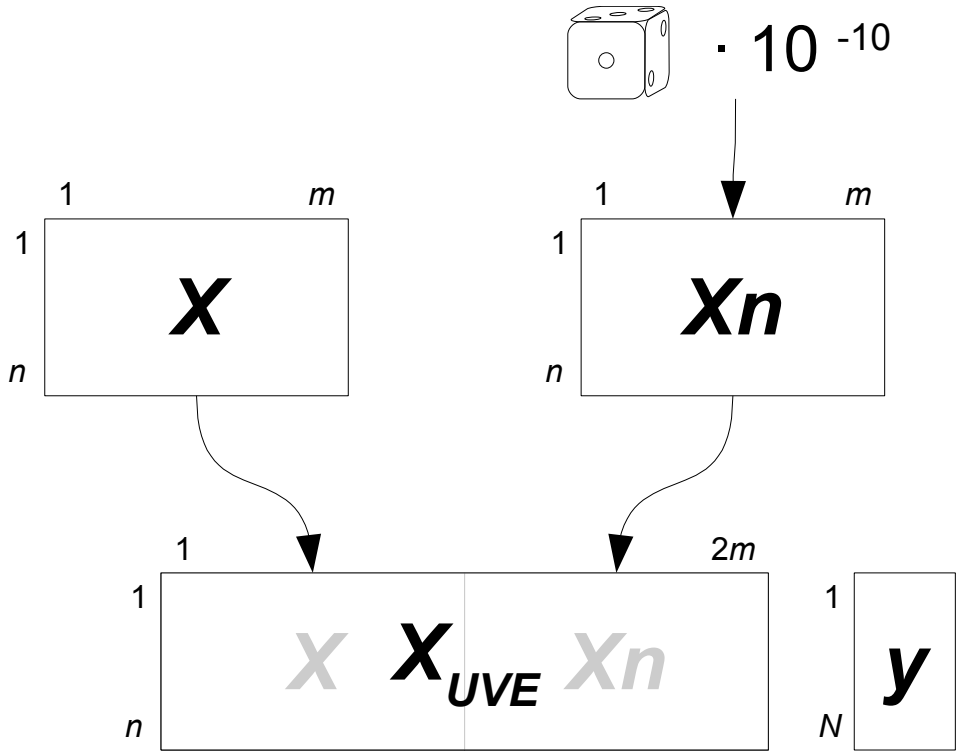
gdzie y_i^{pt} jest elementem i wektora $\boldsymbol{y}^{pt}$ zawierającego prognozowane wartości zmiennej zależnej zbioru testowego, y_i^t jest elementem i wektora $\boldsymbol{y}^t$ zawierającego oryginalne wartości zmiennej zależnej zbioru testowego, n_t jest liczbą obiektów zbioru testowego, $\bar{y}^t$ jest wartością średnią zmiennej zależnej zbioru testowego.

3.6 Wybór / eliminacja zmiennych

Wybór zmiennych jest złożonym problemem modelowania. W przypadku metody PLS zwykle nie ma potrzeby stosowania eliminacji lub wyboru zmiennych jednak specyfika modelowania m-QSAR często wymaga wybrania zmiennych. W literaturze opisanych zostało wiele takich metod. Jedną z nich jest UVE-PLS [73].

3.6.1 Metoda UVE-PLS

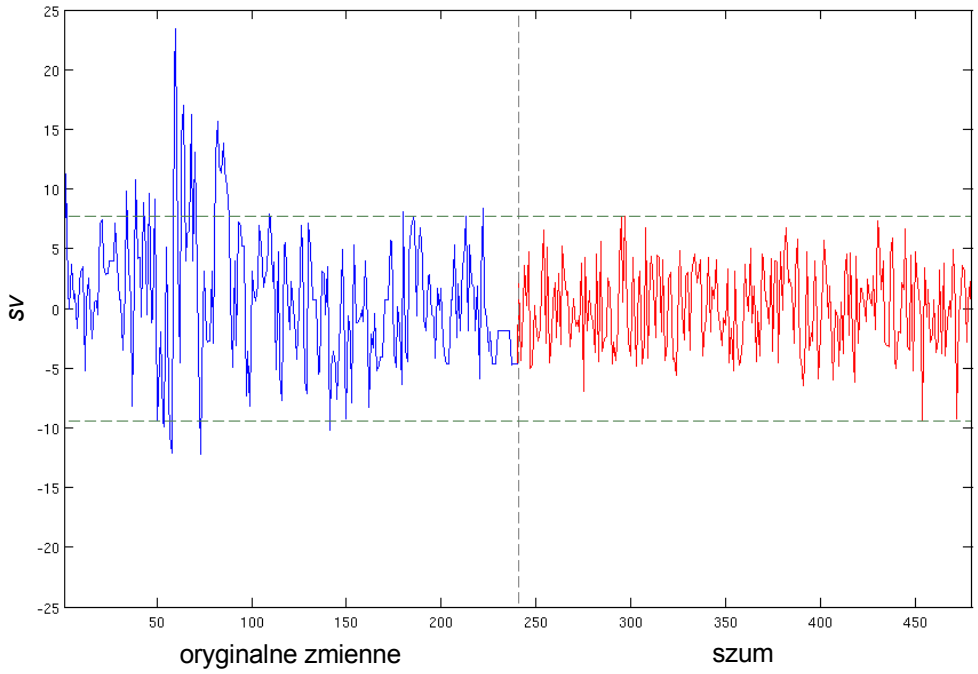
Eliminacja zmiennych metodą UVE-PLS (ang. uninformative variable elimination-PLS) polega na usunięciu zmiennych, które nie niosą istotnej informacji [31, 73, 74]. Kryterium eliminacji jest tzw. stabilność zmiennych (patrz rozdział 3.4.3, strona 34). Macierz oryginalnych zmiennych ***X*** jest uzupełniana o dodatkową macierz szumów ***Xn***. Liczba wierszy macierzy szumu jest równa liczbie wierszy macierzy ***X***, liczba kolumn powinna być porównywalna z liczbą kolumn ***X***. Amplituda szumu musi być niewielka by szum nie wpływał silnie na stabilność oryginalnych zmiennych. Odpowiednia macierz ***Xn*** jest utworzona z wartości losowych pomnożonych przez mały czynnik rzędu 10^{-10} – 10^{-15}. Obie macierze są ze sobą łączone w jedną macierz $\boldsymbol{X}_{UVE}$. Schemat tworzenia macierzy ***Xn*** oraz $\boldsymbol{X}_{UVE}$ jest przedstawiony na rysunku 3.6 [31].



Rysunek 3.6 Macierz ***Xn*** jest tworzona z wartości losowych pomnożonych przez mały czynnik rzędu 10^{-10}. Macierz $\boldsymbol{X}_{UVE}$ jest połączeniem macierzy ***X*** oraz ***Xn***. Symbole *N* oraz *M* ozanaczają odpowiednio liczbę obiektów i liczbę zmiennych macierzy ***X***.

Na podstawie macierzy $\boldsymbol{X}_{UVE}$ oraz wektora ***y*** obliczany jest parametr nazywany stabilnością zmiennych (patrz rozdział 3.4.3, strona 34). Rysunek 3.7 przedstawia wykres stabilności zmiennych macierzy $\boldsymbol{X}_{UVE}$. W części lewej kolorem niebieskim zaznaczono stabilność oryginalnych zmiennych, w części prawej stabilność pochodzącą z macierzy

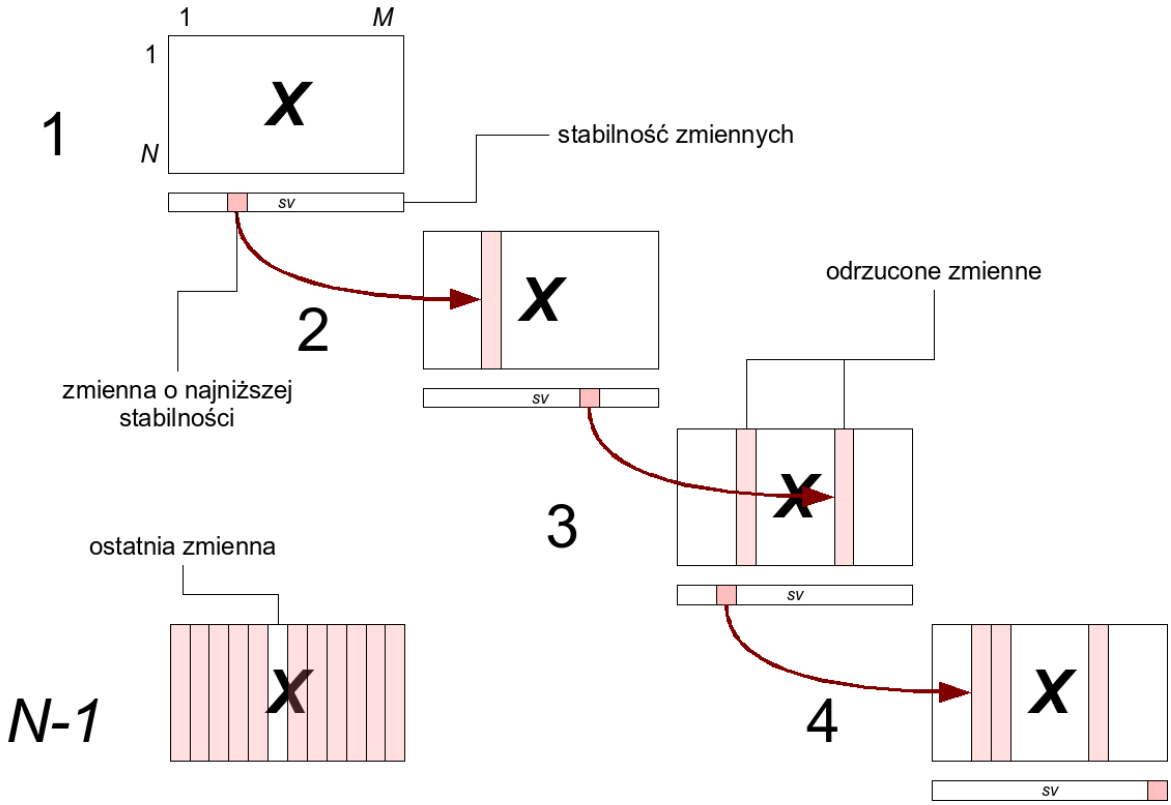
szumów. Działanie metody UVE polega na odrzuceniu zmiennych, których stabilność jest niższa od stabilności szumów. Poziome przerywane linie wskazują zakres stabilności odrzucanych zmiennych.



Rysunek 3.7 Wykres stabilności zmiennych macierzy $\boldsymbol{X}_{UVE}$. Po lewej stronie, kolorem niebieskim, zaznaczono stabilność oryginalnych zmiennych, po prawej stronie, kolorem czerwonym, stabilność dodanych szumów. Poziome przerywane linie wskazują zakres stabilności odrzucanych zmiennych.

3.6.2 Metoda IVE-PLS

Iteracyjna metoda IVE-PLS (ang. iterative variable elimination-PLS) należy do grupy metod wstecznej eliminacji zmiennych (ang. backward elimination). Polega ona na kolejnym odrzucaniu zmiennych o najmniejszym wpływie. Podobnie jak w przypadku UVE-PLS kryterium określającym znaczenie zmiennych jest stabilność. Liczba możliwych iteracji jest o jeden mniejsza od liczby zmiennych. Na rysunku 3.8 przedstawiono schemat blokowy eliminacji IVE-PLS. W kolejnych iteracjach liczona jest stabilność zmiennych. Zmienna o najniższej bezwzględnej stabilności jest eliminowana z macierzy $\boldsymbol{X}$ (wyeliminowane zmienne znaczono kolorem jasnoczerwonym) [31].

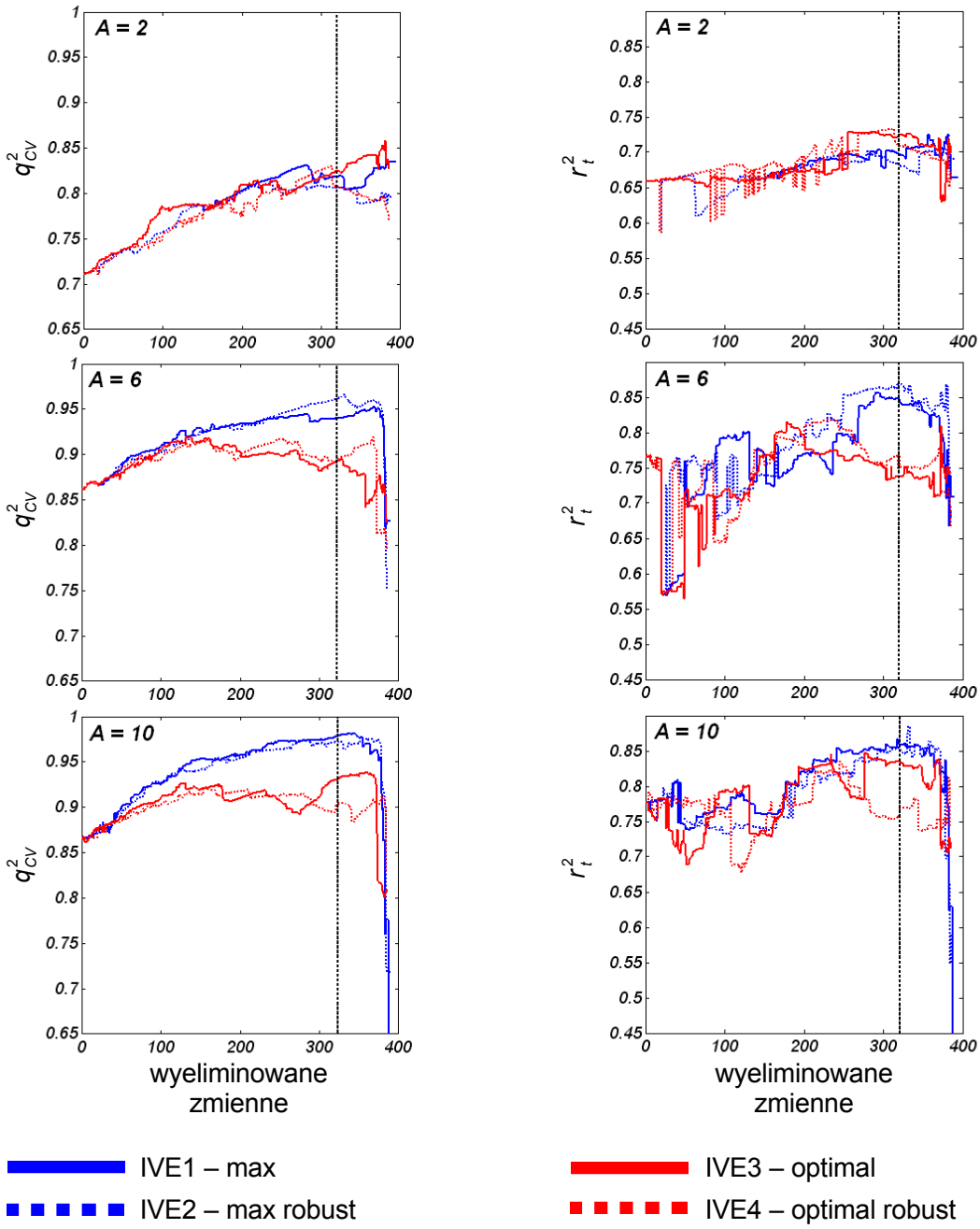


Rysunek 3.8 Schemat eliminacji zmiennych metodą IVE-PLS. W kolejnych iteracjach 1, 2, 3, 4, ..., *N*-2 odrzucane są zmienne o najniższej stabilności. W ostatniej iteracji *N*-1 zostaje jedna zmienna.

Podczas całej procedury w każdej iteracji obliczane są parametry testujące jakość modelu. Zwykle obliczane są parametry q^2_{CV} oraz r^2_t o ile jest dostępny zewnętrzny zbiór testowy. Badanie przebiegu q^2_{CV} w czasie całej procedury pozwala wybrać optymalny zestaw zmiennych. Rysunek 3.9 przedstawia zmiany parametrów q^2_{CV} i r^2_t podczas modelowania SOM-CoMSA szeregu pochodnych kwasu karboksylowego przy użyciu czterech różnych procedur eliminacji IVE-PLS dla różnych założonych kompleksowości modelu [84]. Poszczególne procedury eliminacji różniły się od siebie sposobem obliczania stabilności. W procedurach IVE1 oraz IVE3 zastosowano stabilności standardowe – wzór (3.15). W procedurach IVE2 oraz IVE4 zastosowano elastyczną stabilność (robust) – wzór (3.16). Co więcej, w przypadku procedur IVE1 oraz IVE2 stabilność była obliczona dla maksymalnej założonej kompleksowości (max). Natomiast w przypadku procedur IVE3 oraz IVE4 stabilność była obliczana dla kompleksowości optymalnej (optimal).

Wartości q^2_{CV} w kolejnych procedurach przez pierwsze 100 iteracji różnią się od siebie tylko nieznacznie. Po około 100 iteracjach wartości q^2_{CV} uzyskane dla procedur IVE1/IVE2 przewyższają uzyskane dla procedur IVE3/IVE4. Jest to zwłaszcza wyraźne w przypadku wyższych kompleksowości. Wartości r^2_t wskazują na wysoką zdolność prognozowania modeli aczkolwiek na wykresach widać znaczną niestabilność wartości parametru.

Wybór optymalnej liczby iteracji (czyli optymalnego zestawu zmiennych) może być wykonany na wiele sposobów. Najprostszą metodą jest wybór punktu, w którym q^2_{CV} osiąga wartość maksymalną. Często eliminację prowadzi się do z góry ustalonego punktu. Na przykład, na rysunku 3.9 pionowa przerywana linia oznacza 80% odrzuconych zmiennych. Użycie tak określonego punktu umożliwia łatwe porównywanie wyników różnych eliminacji. W rozdziałach 7.1.1 (strona 81) oraz 7.1.2 (strona 85) znajduje się omówienie zastosowań IVE-PLS w modelowaniu s-CoMSA.



Rysunek 3.9 Przebiegi parametrów q^2_{CV} oraz r^2_t uzyskane w wyniku modelowania CoMSA szeregu pochodnych kwasu karboksylowego przy użyciu różnych procedur eliminacji IVE-PLS dla *a priori* założonych maksymalnych kompleksowości modelu, odpowiednio A = 2, 6 oraz 10. Według [84].

4 Wizualizacja modeli 3D-QSAR

Celem modelowania QSAR jest ilościowe powiązanie struktur związków chemicznych przedstawionych w postaci deskryptora molekularnego z aktywnością. Uzyskana funkcyjna zależność może być użyta do przewidywania aktywności nowych związków. W przypadku wykorzystania jako zmiennych tzw. zmiennych ukrytych nie tłumaczy ona jednak w prosty sposób molekularnych podstaw badanego efektu.

Wykorzystanie metod QSAR do ustalania molekularnych uwarunkowań aktywności opiera się przede wszystkim na wizualizacji modeli. Poprzez wizualizację modeli rozumie się obrazowanie w trójwymiarowej przestrzeni składowych deskryptora molekularnego o największym wkładzie w modelowany efekt. Składowe deskryptora są wyświetlane razem z reprezentatywną cząsteczką analizowanego szeregu związków chemicznych, dzięki czemu uwidaczniają się elementy struktury powiązane z wyświetlonymi składowymi deskryptora.

Identyfikacja elementów deskryptora mających największy wkład w modelowany efekt może być dokonana na wiele różnych sposobów. W tym celu bada się wariancję poszczególnych zmiennych, ich korelację z aktywnością, stosuje się wartości progowe oraz metody wyboru i eliminacji zmiennych – patrz również rozdziały 7.1.2 (strona 85) oraz 7.2 (strona 91).

4.1 Mapy konturowe

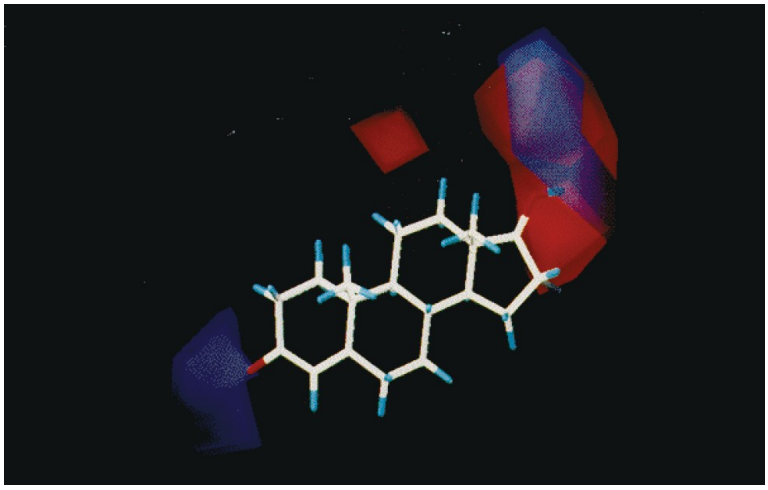
Metody QSAR wykorzystujące koncepcję pól molekularnych, tj. metoda CoMFA, CoMSIA i inne podobne, najczęściej wykorzystują do wizualizacji modeli tzw. mapy konturowe (ang. contour maps). Wykres tego typu powstaje przez odrzucenie punktów pola, których wartości nie przekraczają ustalonych progów. Zwykle stosowane są dwie wartości progowe: górna i dolna – odrzucane są więc punkty, których wartości mieszczą się pomiędzy progami. Natomiast pozostałe punkty wskazują obszary przestrzeni wykazujące odpowiednio pożądane i niepożądane oddziaływania z cząsteczkami szeregu.

Rodzaj wyświetlanych oddziaływań zależy od wybranego pola molekularnego. Standardowo w metodzie CoMFA używane są dwa rodzaje pól – pole oddziaływań elektrostatycznych oraz pole oddziaływań sterycznych. W metodzie CoMSIA używane są

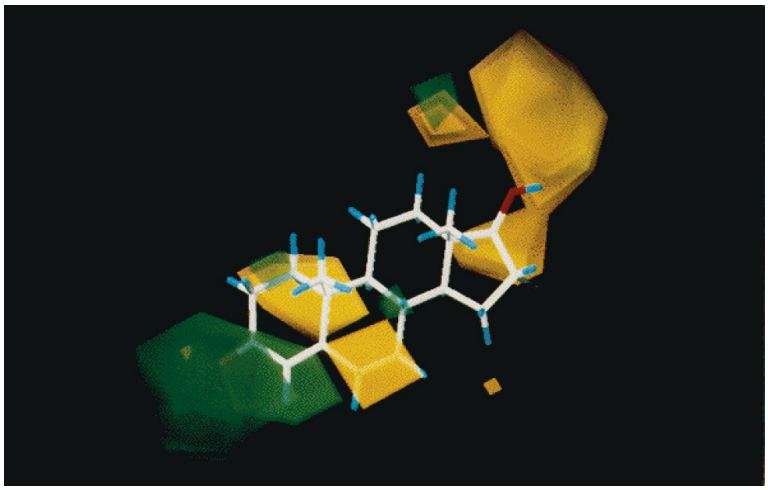
dodatkowo pola oddziaływań hydrofobowych oraz dwa pola oddziaływań wodorowych, jedno wskazuje na występowanie donorów, drugie na występowanie akceptorów wiązań wodorowych.

Obok typu pola istotne jest wybranie rodzaju transformacji wartości pól (ang. type of transformation). W metodzie CoMFA domyślna transformacja polega na przemnożeniu odchylenia standardowego punktów pola przez współczynniki regresji. W literaturze transformacja ta oznaczana jest symbolem `StDev*Coeff`. Stosowane są również inne transformacje np. `Mean*Coeff` (wartość średnia pomnożona przez współczynniki regresji), `Average_Field` (wartość średnia) itp. [13].

a)



b)



Rysunek 4.1 Mapy konturowe uzyskane w wyniku analizy CoMFA steroidów o powinowactwie TBG [85]. Szczegóły w tekście.

Wartości progów są ustalane na dwa sposoby. Są to albo wartości rzeczywiste wyrażone w kcal/mol albo wartości procentowe. Standardowo używane są wartości procentowe. W przypadku progów procentowych górna wartość przyjmuje zwykle 80% a dolna 20%.

Wybór odpowiedniego typu pola, transformacji i wartości progowych pozwala uzyskać mapy konturowe wskazujące obszary o pożądanych i niepożądanych obszarach oddziaływań elektrostatycznych, sterycznych i innych. Rysunek 4.1 przedstawia przykładowe mapy konturowe. Rysunek zaczerpnięto z publikacji [85]. Kolor niebieski oznacza obszary, w których obecność podstawnika elektroujemnego zwiększyłaby aktywność a kolor czerwony, w których taki podstawnik zmniejszyłby aktywność. Kolor zielony natomiast wskazuje obszary w których pożądane są oddziaływania steryczne a kolor żółty wskazuje obszary, w których takie oddziaływania nie są pożądane.

4.2 Wizualizacja oddziaływań specyficznych

Deskryptory molekularne stosowane w metodzie CoMFA oraz w metodach pokrewnych nie mogą być bezpośrednio powiązane z określonymi fragmentami struktury związków chemicznych. Dlatego do wizualizacji modeli QSAR uzyskanych tymi metodami stosuje się mapy konturowe, wskazujące jedynie pewne obszary przestrzeni o pożądanych bądź niepożądanych oddziaływaniach.

W przypadku metod QSAR wykorzystujących deskryptory specyficzne, tj. deskryptory, które można jednoznacznie powiązać z określonymi fragmentami struktury, wizualizacja modeli umożliwia precyzyjne zaznaczenie fragmentów struktury mających największy wkład w modelowany efekt. Ponieważ efekt modelowany metodami QSAR jest w ogromnej większości przypadków powiązany z oddziaływaniem cząsteczek z receptorem dlatego można przypuszczać, że fragmenty struktury wskazane poprzez wizualizację są odpowiedzialne za specyficzne oddziaływania z receptorem.

5 Badania własne

Cząsteczka w metodzie CoMFA (oraz metodach pokrewnych) reprezentowana jest poprzez deskryptor złożony z szeregu liczb definiujących pole cząsteczkowe w równomiernie rozłożonej sieci przestrzennej. Jest to więc wielopunktowa reprezentacja cząsteczek.

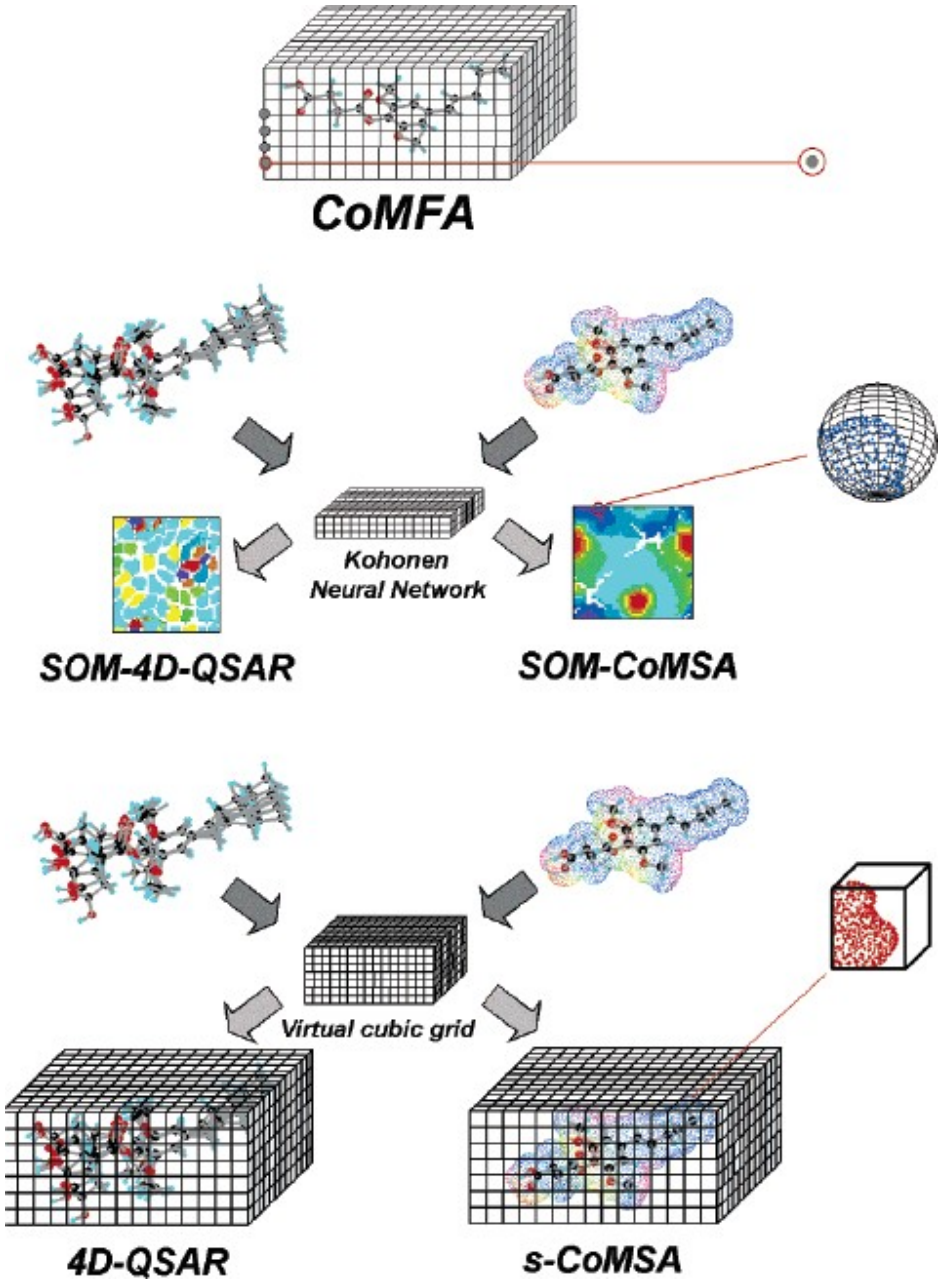
Deskryptory molekularne można również obliczyć poprzez podział molekuł na części. Uzyskane w wyniku takiego podziału przestrzenne sektory mogą zawierać pojedyncze atomy, grupy atomów lub mogą być puste. Podział cząsteczek na przestrzenne sektory nazywany jest formalizmem sektorowym. Pierwszy raz taki podział zaproponowali Purcell i Testa [86].

Metoda 4D-QSAR Hopfingera używa formalizmu sektorowego do opisu przestrzeni konformacyjnej analizowanych molekuł [48]. Formalizm sektorowy jest również stosowany w opisywanej w tej pracy sektorowej wersji metody CoMSA (s-CoMSA). Ciekawym sposobem podziału cząsteczek na przestrzenne fragmenty jest zastosowanie sieci neuronowej Kohonena w neuronowej wersji metody CoMSA (SOM-CoMSA). Każdy neuron w tej metodzie obejmuje obszar w przestrzeni co powoduje podział powierzchni cząsteczkowej na fragmenty. Sieci neuronowe Kohonena są również stosowane w metodzie SOM-4D-QSAR do podziału na części przestrzeni konformacyjnej analizowanych cząsteczek [87].

Rysunek 5.1 przedstawia schemat ilustrujący różnice między reprezentacją cząsteczek stosowaną w metodzie CoMFA oraz w metodach wykorzystujących formalizm sektorowy oraz sieci neuronowej Kohonena tj. w metodach s-CoMSA, 4D-QSAR, SOM-CoMSA, oraz SOM-4D-QSAR.

W metodzie CoMFA generowanie deskryptora polega na obliczaniu wartości odpowiedniego pola cząsteczkowego w ściśle określonych punktach przestrzeni (w węzłach siatki). Natomiast w metodach 4D-QSAR, s-CoMSA, SOM-CoMSA i SOM-4D-QSAR zamiast pojedynczych punktów definiuje się pewne zakresy przestrzeni. Generowanie deskryptora w tych metodach polega na obliczeniu odpowiednich własności w zdefiniowanych zakresach przestrzeni. W przypadku metod 4D-QSAR i s-CoMSA w

miejsce punktów stosowane są sektory. Metody SOM-4D-QSAR i SOM-CoMSA stosują formalizm samoorganizujących się map neuronowych, których działanie sprowadza się do definiowania sfer obejmujących pewien zakres przestrzeni.



Rysunek 5.1 Schemat obrazujący różnice między reprezentacją cząsteczek stosowaną w metodzie CoMFA oraz w metodach s-CoMSA, 4D-QSAR, SOM-CoMSA i SOM-4D-QSAR [29].

5.1 Sektorowa porównawcza analiza powierzchni cząsteczkowych

Metoda sektorowej porównawczej analizy powierzchni cząsteczkowych s-CoMSA (ang. sector-comparative molecular surface analysis) jest metodą 3D-QSAR. W metodzie tej cząsteczka reprezentowana jest przez powierzchnię cząsteczkową. Analogicznie jak w metodzie SOM-CoMSA dla punktów pobieranych z takiej powierzchni oblicza się wartość potencjału elektrostatycznego lub innej własności molekularnej [27]. Następnie przestrzeń dzieli się na jednostkowe sektory (sześciany). Obliczone średnie lub sumowane wartości odpowiedniego deskryptora w sektorach porządkuje się w postaci wektorów opisujących cząsteczki.

Przebieg analizy QSAR metodą s-CoMSA można podzielić na następujące etapy:

- Modelowanie trójwymiarowych obrazów cząsteczek

Przed przystąpieniem do modelowania 3D-QSAR konieczne jest wygenerowanie trójwymiarowych struktur analizowanych cząsteczek. Obliczane są również cząstkowe ładunki atomowe potrzebne do generowania potencjału elektrostatycznego.

- Superpozycja obrazów 3D

Deskryptor metody s-CoMSA jest zależny od wzajemnego położenia analizowanych cząsteczek, dlatego konieczne jest ich nałożenie na wspólny wzorzec. Przez wzorzec nakładania rozumie się motyw strukturalny wspólny dla wszystkich cząsteczek. Najczęściej jako wzorzec nakładania stosuje się najbardziej aktywną cząsteczkę analizowanego szeregu [88, 89, 90]. Jeżeli istnieje hipoteza farmakoforowa powiązana z modelowanym efektem powinna być ona uwzględniona w procesie nakładania [14].

- Generowanie powierzchni cząsteczkowych, obliczanie potencjałów

Powierzchnie cząsteczkowe są generowane po etapie nakładania, gdy jest już ustalona geometria i wzajemne położenie w przestrzeni. Metoda s-CoMSA wymaga próbkowania powierzchni punktami opisanymi współrzędnymi trójwymiarowej przestrzeni. Dodatkowo dla każdego punktu obliczana jest wartość wybranego potencjału. Tak więc każdy punkt opisany jest wektorem $\boldsymbol{d}(x, y, z, v)$. Standardowo oblicza się potencjał elektrostatyczny, aczkolwiek może to być inna własność, np. potencjał lipofilowy. Wybór potencjału jest uzależniony od rodzaju modelowanego efektu.

- Generowanie wirtualnej siatki

Wymiary siatki są ustalane w ten sposób, by siatka obejmowała powierzchnie wszystkich analizowanych cząsteczek. Dodatkowo siatka jest poszerzana z każdej strony o ustalony margines, zwykle około 1 Å. Następnie siatka dzielona jest na sześcienne sektory o określonym rozmiarze, najczęściej o szerokości od 1 do 2 Å.

- Obliczanie deskryptora s-CoMSA

Podział powierzchni cząsteczkowych polega na przypisaniu punktów próbkowanych na powierzchni odpowiednim sektorom. Dla każdej cząsteczki tworzony jest wierszowy wektor o długości równej liczbie sektorów. Każdy element wektora odpowiada jednemu sektorowi. Wartości elementów wektora są obliczane na podstawie punktów powierzchni obejmowanych przez odpowiadające im sektory. Wektory wszystkich analizowanych cząsteczek są następnie łączone w macierz charakteryzującą cały szereg – macierz $\boldsymbol{X}$.

- Modelowanie i walidacja

Etap modelowania i walidacji prowadzi się metodą PLS.

- Wizualizacja modeli

Eliminacja zmiennych, np. metodą IVE, pozwala znacznie ograniczyć liczbę zmiennych, pozostawiając tylko te, które mają największy wkład w modelowanie aktywności. Wizualizacja obszarów powierzchni cząsteczkowych odpowiadających tym zmiennym daje obraz prawdopodobnych obszarów oddziaływań cząsteczki i receptora. Przez receptor, w tym kontekście, rozumie się cel molekularny mający bezpośrednie powiązanie z modelowaną aktywnością. Każdemu zaznaczonemu obszarowi można dodatkowo przypisać wagę z jaką wchodzi do modelu. Analiza uzyskanych obrazów oddziaływań specyficznych pozwala zrozumieć relacje miedzy aktywnością a strukturą analizowanych cząsteczek chemicznych.

5.2 Formalizm metody s-CoMSA

Formalnie metoda s-CoMSA obejmuje transformację powierzchni cząsteczkowych w wektory o określonej długości – deskryptory s-CoMSA. Wektor uzyskany w wyniku takiej transformacji w sposób liczbowy opisuje powierzchnię cząsteczki. Długość wektora i sens jego składowych zależy od parametrów transformacji. Zbiór takich parametrów generuje dla danego szeregu cząsteczkowego wektory o jednakowej długości.

Każda cząsteczka poddawana transformacji s-CoMSA jest reprezentowana przez powierzchnię. Niech P_m będzie powierzchnią cząsteczki m a p_{mj} niech będzie elementem j zbioru P_m (tj. punktem j powierzchni P_m). Każdy punkt powierzchni opisany jest co najmniej czterema współrzędnymi $p_{mj}(x)$, $p_{mj}(y)$, $p_{mj}(z)$, $p_{mj}(v)$ gdzie x, y, z oznaczają współrzędne w trójwymiarowej przestrzeni a v oznacza własność punktu powierzchni, np. wartość potencjału elektrostatycznego.

Niech S będzie zbiorem wszystkich sektorów. Elementami zbioru S są sektory s_i. Każdy sektor s_i definiowany jest w układzie współrzędnych przez konstrukcję par płaszczyzn równoległych do głównych płaszczyzn układu współrzędnych wyznaczonych odpowiednio przez osie układu. Dla współrzędnych x, y, z są to płaszczyzny YZ, XZ, XY. Każda para płaszczyzn przecina odpowiednią oś układu w dwóch punktach określających dolny i górny kres odpowiednich zmiennych x, y, z wyznaczających przestrzeń danego sektora. Niech wektory $\boldsymbol{l_i}$ oraz $\boldsymbol{h_i}$ wyznaczają odpowiednio kresy dolne i górne przestrzeni obejmowanej przez sektor s_i na kolejnych osiach układu współrzędnych.

Niech K_{mi} będzie podzbiorem zbioru P_m takim, że wszystkie jego elementy są zawarte w sektorze s_i. Punkt p_{mj} należy do zbioru K_{mi} wtedy i tylko wtedy gdy j spełnia warunek:

$$\begin{array}{lllllll} p_{mj}(x) & \geq & l_i(x) & \wedge & p_{mj}(x) & < & h_i(x) \quad \wedge \\ p_{mj}(y) & \geq & l_i(y) & \wedge & p_{mj}(x) & < & h_i(x) \quad \wedge \\ p_{mj}(z) & \geq & l_i(z) & \wedge & p_{mj}(x) & < & h_i(x) \end{array} \tag{5.1}$$

Jeżeli dla żadnego j powyższy warunek nie jest spełniony zbiór K_{mi} jest zbiorem pustym.

Podział na wirtualne sektory przeprowadza się w taki sposób by wyznaczone jednostkowe sektory posiadały jednakową objętość i kształt. W opisywanej metodzie spełniony jest również warunek rozłączności sektorów:

$$\underset{i,\,i'}{\forall}\; i \neq i' \;\; s_i \cap s_{i'} = \emptyset \tag{5.2}$$

W ogólnym przypadku warunek (5.2) ogranicza redundancję danych i nie musi być spełniony.

Niech $\boldsymbol{w}$ będzie wektorem deskryptora s-CoMSA a w_i niech będą jego składowymi. Wartości składowych w_i są obliczane za pomocą funkcji C( ) działającej na zbiorze K_{mi}. Standardowo dla wszystkich składowych wektora $\boldsymbol{w}$ stosowana jest ta sama funkcja

transformująca. Zastosowanie różnych funkcji C( ) do obliczania poszczególnych składowych wektora $\boldsymbol{w}$ wiązałoby się z formalną koniecznością uzupełnienia definicji sektorów s_i o właściwą dla nich funkcję transformującą $C_i( )$.

Zdefiniowano trzy funkcje transformujące. Funkcja oznaczona symbolem *mvp* oblicza średnią wartość potencjału punktów zbioru K_{mi}:

$$C_{mvp}(K_{mi}) = \begin{cases} \dfrac{\sum_f k_{mif}(v)}{|K_{mi}|} & , \text{gdy } K_{mi} \neq \emptyset \\ 0 & , \text{gdy } K_{mi} = \emptyset \end{cases} \tag{5.3}$$

gdzie k_{mif} jest elementem f zbioru K_{mi}. Funkcja oznaczona symbolem *nps* oblicza względną gęstość punktów zawartych w zbiorze K_{mi}:

$$C_{nps}(K_{mi}) = \begin{cases} \dfrac{|K_{mi}|}{d_m} & , \text{gdy } K_{mi} \neq \emptyset \\ 0 & , \text{gdy } K_{mi} = \emptyset \end{cases} \tag{5.4}$$

gdzie d_m jest wartością stałą dla powierzchni m, a symbol $|K_{mi}|$ oznacza moc zbioru. Stała d_m zapewnia normalizację wyniku dla powierzchni o różnej gęstości punktów.
Funkcja oznaczona symbolem *mvp/nps* jest niejako ilorazem funkcji *mvp* oraz *nps*:

$$C_{mvp/nps}(K_{mi}) = \begin{cases} \dfrac{\sum_f k_{mif}(v)}{|K_{mi}|^2} d_m & , \text{gdy } K_{mi} \neq \emptyset \\ 0 & , \text{gdy } K_{mi} = \emptyset \end{cases} \tag{5.5}$$

Metoda s-CoMSA, w ujęciu matematycznym, sprowadza się do generowania macierzy $\boldsymbol{X}$ o wymiarach $g \times n$, gdzie g jest liczbą analizowanych cząsteczek a n jest liczbą sektorów czyli długością wektora $\boldsymbol{w}$. Element macierzy $\boldsymbol{X}$, x_{mi} definiowany jest następującym wzorem:

$$x_{mi} = C(K_{mi}) \tag{5.6}$$

gdzie x_{mi} jest elementem macierzy $\boldsymbol{X}$.

Macierz $\boldsymbol{X}$ wyznaczona za pomocą wzoru (5.6) bez odwoływania się do żadnych obiektów referencyjnych jest oznaczana symbolem $\boldsymbol{X}_A$. W analogii do metody 4D-QSAR deskryptor zawarty w takiej macierzy został nazwany absolutnym (ang. absolute). Natomiast zdefiniowanie zbioru cząsteczek referencyjnych w postaci macierzy $\boldsymbol{X}^R$, obliczonej ze wzoru (5.6), pozwala wyznaczyć deskryptory typu łącznego (ang. joint) oraz rozłącznego (ang. self) w postaci odpowiednio macierzy $\boldsymbol{X}_J$, $\boldsymbol{X}_S$. Metodyka s-CoMSA umożliwia użycie dowolnej liczby cząsteczek referencyjnych jednakże standardowo stosowana jest tylko jedna. Macierz $\boldsymbol{X}^R$ staje się wówczas wektorem wierszowym.

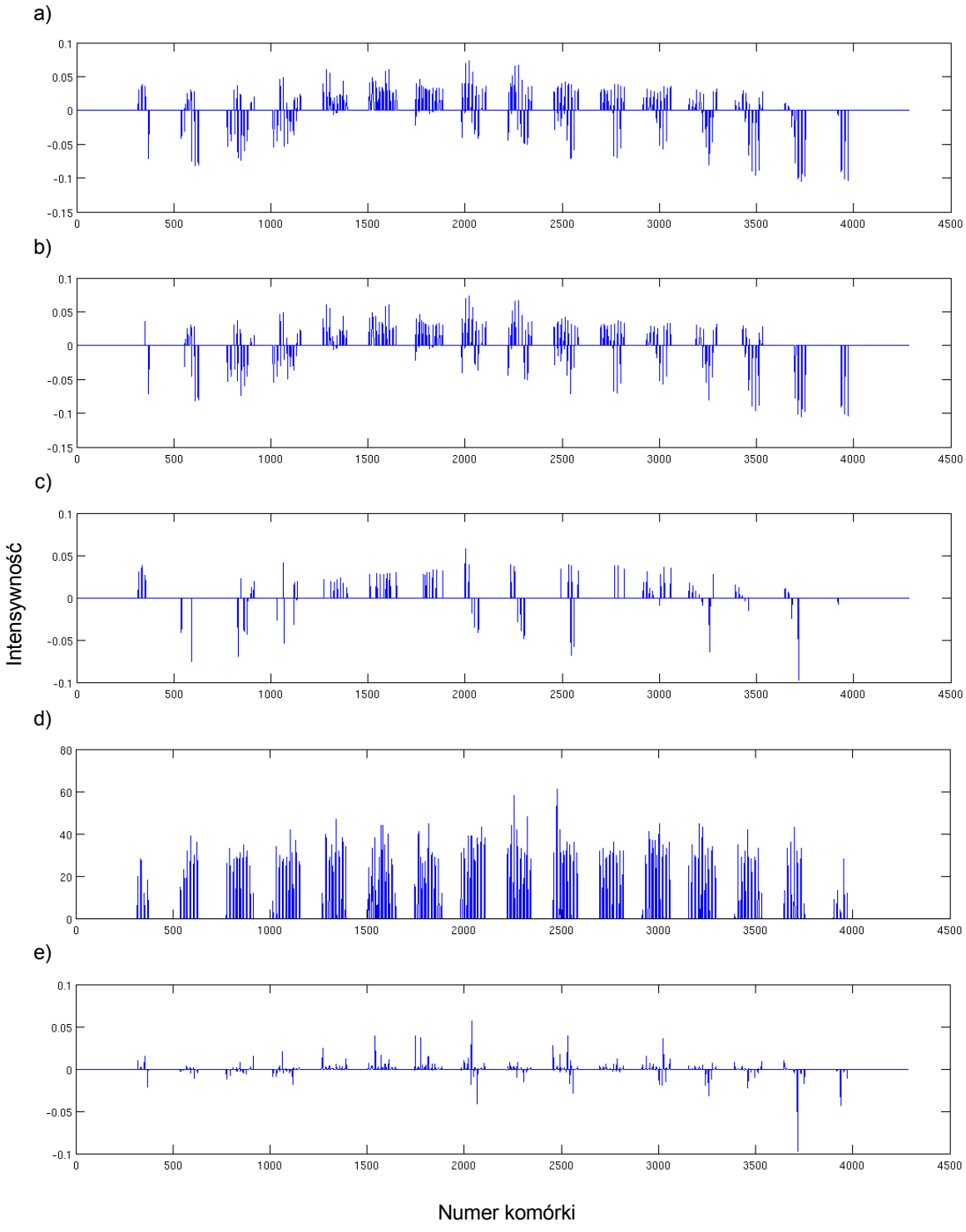
Elementy macierzy deskryptora łącznego $\boldsymbol{X}_J$ oraz macierzy deskryptora rozłącznego $\boldsymbol{X}_S$, $x_{J\,mi}$, oraz $x_{S\,mi}$, definiowane są odpowiednio wzorami uwzględniającymi macierz cząsteczek referencyjnych:

$$x_{J\,mi}=\begin{cases} \mathrm{C}\left(K_{mi}\right) & \text{, gdy } K_{mi}\neq\emptyset \wedge \sum\limits_{m^R}\left(x^R_{m^R i}\right)^2 \neq 0 \\ 0 & \text{, gdy } K_{mi}=\emptyset \vee \sum\limits_{m^R}\left(x^R_{m^R i}\right)^2 = 0 \end{cases} \tag{5.7}$$

$$x_{S\,mi}=\begin{cases} \mathrm{C}\left(K_{mi}\right) & \text{, gdy } K_{mi}\neq\emptyset \wedge \sum\limits_{m^R}\left(x^R_{m^R i}\right)^2 = 0 \\ 0 & \text{, gdy } K_{mi}=\emptyset \vee \sum\limits_{m^R}\left(x^R_{m^R i}\right)^2 \neq 0 \end{cases} \tag{5.8}$$

gdzie indeks m^R dotyczy cząsteczek referencyjnych a $x^R_{m^R i}$ jest elementem macierzy $\boldsymbol{X}^R$.

Rysunek 5.2 przedstawia wykresy MSS (ang. molecular shape spectrum) różnych deskryptorów s-CoMSA. Pierwsze trzy wykresy przedstawiają wykresy MSS deskryptorów typu absolutnego, łącznego i rozłącznego uzyskane dla standardowej funkcji transformującej *mvp*. Dwa ostatnie wykresy przedstawiają deskryptory typu absolutnego uzyskane dla funkcji transformującej *nps* oraz *mvp/nps*.



Rysunek 5.2 Wykresy MSS związku **s31** (patrz rozdział 5.4.1.1, strona 52): deskryptor typu absolutnego (a), łącznego (b) i rozłącznego (c) uzyskany dla funkcji transformującej *mvp*; deskryptory typu absolutnego uzyskane dla funkcji transformującej *nps* (d) oraz *mvp/nps* (e). Do wykonania wykresów (b) oraz (c) użyto związku referencyjnego **s6**. (patrz również rozdział 5.4.1.1, strona 52).

5.3 Rozmiar komórki siatki – gęstość siatki

Podstawowym parametrem charakteryzującym deskryptor s-CoMSA jest rozmiar komórki wirtualnej siatki. Rozmiar komórki ma istotny wpływ na rozdzielczość reprezentacji trójwymiarowych powierzchni cząsteczkowych (patrz rozdział 5.4.2, strona 59). Wirtualna siatka jest tworzona w taki sposób, że wszystkie sektory (komórki siatki) są sześcianami o jednakowych rozmiarach. Jako miarę rozmiaru komórki przyjęto więc długość krawędzi sektora i oznaczono symbolem *cs*. Formalnie rozmiar komórki jest zdefiniowany za pomocą wzoru:

$$cs = \frac{\boldsymbol{h}(x) - \boldsymbol{l}(x) + \boldsymbol{h}(y) - \boldsymbol{l}(y) + \boldsymbol{h}(z) - \boldsymbol{l}(z)}{3} \tag{5.9}$$

gdzie $\boldsymbol{h}$ oraz $\boldsymbol{l}$ są wektorami wyznaczającymi kres górny i dolny sektora na poszczególnych osiach układu współrzędnych.

Formalizm metody s-CoMSA umożliwia użycie sektorów o różnych długościach krawędzi dlatego parametr *cs* jest obliczany ze wzoru (5.9) uwzględniającego wszystkie krawędzie. W przypadku standardowej analizy s-CoMSA sektory są sześcianami foremnymi.

5.4 Testowanie metody s-CoMSA

5.4.1 Modelowane efekty i szeregi molekularne

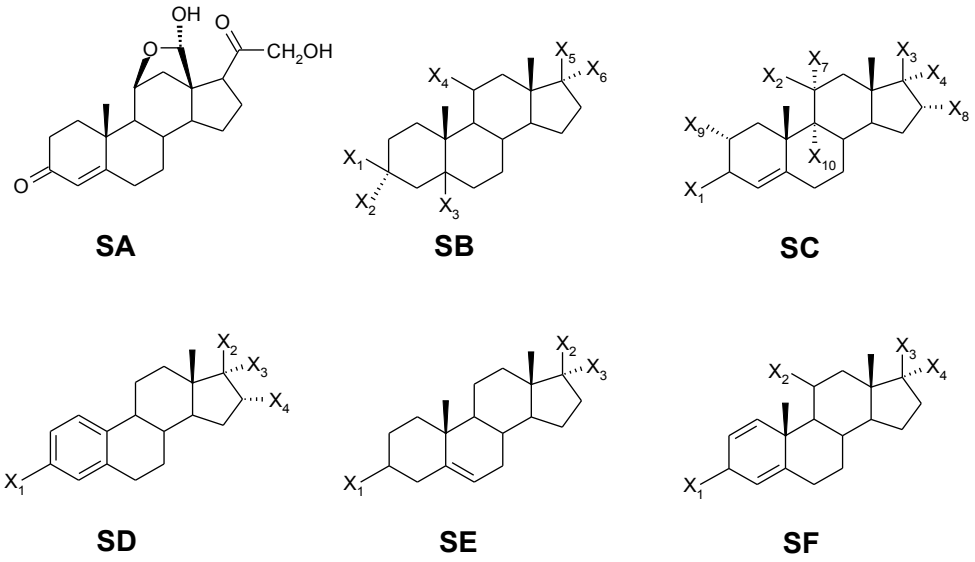
5.4.1.1 Szereg steroidów o powinowactwie do globuliny wiążącej kortykosteroidy

Ze względu na sztywność szkieletu steroidowego szereg steroidowych pochodnych wykazujących powinowactwo do globuliny wiążącej kortykosteroidy (ang. corticosteroid-binding globulin – CBG) wykorzystywany jest do rutynowego testowania różnych metod 3D-QSAR [21]. Wygenerowane trójwymiarowe struktury zostały nałożone wszystkimi atomami węgla tworzącymi czteropierścieniowy szkielet steroidowy. Nakładanie było wykonane za pomocą programu Match3D i polegało na zminimalizowaniu odległości między nakładanymi atomami a atomami wzorca [91]. Podobnie jak w innych analizach 3D-QSAR tego szeregu jako wzorzec nakładania wybrano pochodną **s6**. Jest to jeden z dwóch najaktywniejszych związków analizowanego szeregu. Struktury steroidów, oraz ich

powinowactwo do globuliny wiążącej kortykosteroidy przedstawiono w tabeli 5.1, natomiast struktury sześciu szkieletów steroidowych występujących w analizowanej serii przedstawiono na rysunku 5.3.

Podobnie jak w innych pracach szereg został podzielony na zbiór modelowy, związki od **s1** do **s21** oraz na zbiór testowy, związki **s22** do **s31** [92]. Związki **s1**-**s21** (zbiór modelowy) wykorzystano do modelowania PLS aktywności CBG. Związki **s22**-**s31** (zbiór testowy) służył do walidacji zdolności prognozowania modeli generowanych w oparciu o zbiór modelowy. Maksymalne wartości q^2_{CV} wahały się w zależności od modelu w granicach 0,88 – 0,90 a wartości *SDEP* w granicach 0,78 – 0,73.

Uzyskane wartości są porównywalne lub lepsze od wartości uzyskanych innymi metodami: CoMFA q^2_{CV} = 0,73 [92], *SDEP* = 0,84 [20]; SOM-CoMSA q^2_{CV} = 0,88, *SDEP* = 0,69 [21]; Quasar q^2_{CV} = 0,90 [54].



Rysunek 5.3 Typy szkieletów steroidowych występujące w grupie 31 analizowanych steroidów CBG. Podpisy pod kolejnymi szkieletami odpowiadają kolumnie **S** w tabeli 5.1, symbole grup funkcyjnych od X_1 do X_{10} odpowiadają odpowiednio kolumnom od $\mathbf{X_1}$ do $\mathbf{X_{10}}$ w tej samej tabeli.

Tabela 5.1 Aktywności CBG i struktury analizowanych steroidów. Kolumna **S** oznacza typ szkieletu steroidowego, kolumny $\mathbf{X_1}$ do $\mathbf{X_{10}}$ odpowiadają grupom funkcyjnym opisanym na rysunku 5.3.

Nr	S	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	CBG
s1	SA											-6,279
s2	SB	OH	H	H[a]	H	OH	H					-5,000
s3	SE	OH	OH	H								-5,000
s4	SC	=O	H	=O				H	H	H	H	-5,763
s5	SB	H	OH	H[a]	H	=O						-5,613
s6	SC	=O	OH	$COCH_2OH$	H			H	H	H	H	-7,881
s7	SC	=O	OH	$COCH_2OH$	OH			H	H	H	H	-7,881
s8	SC	=O	=O	$COCH_2OH$	OH				H	H	H	-6,892
s9	SE	OH	=O									-5,000
s10	SC	=O	H	$COCH_2OH$	H			H	H	H	H	-7,653
s11	SC	=O	H	$COCH_2OH$	OH			H	H	H	H	-7,881
s12	SB	=O		H[a]	H	OH	H					-5,919
s13	SD	OH	OH	H	H							-5,000
s14	SD	OH	OH	H	OH							-5,000
s15	SD	OH	=O		H							-5,000
s16	SB	H	OH	H[b]	H	=O						-5,255
s17	SE	OH	COMe	H								-5,255
s18	SE	OH	COMe	OH								-5,000
s19	SC	=O	H	COMe	H			H	H	H	H	-7,380
s20	SC	=O	H	COMe	OH			H	H	H	H	-7,740
s21	SC	=O	H	OH	H			H	H	H	H	-6,724
s22	SF	=O	OH	$COCH_2OH$	OH							-7,512
s23	SC	=O	OH	$COCH_2OCOMe$	OH			H	H	H	H	-7,553
s24	SC	=O	=O	COMe	H				H	H	H	-6,779
s25	SC	=O	H	$COCH_2OH$	H			OH	H	H	H	-7,200
s26	SC[c]	=O	H	OH	H			H	H	H	H	-6,144
s27	SC	=O	H	COMe	OH			H	OH	H	H	-6,247
s28	SC	=O	H	COMe	H			H	Me	H	H	-7,120
s29	SC[c]	=O	H	COMe	H			H	H	H	H	-6,817
s30	SC	=O	OH	$COCH_2OH$	OH			H	H	Me	H	-7,688
s31	SC	=O	OH	$COCH_2OH$	OH			H	H	Me	F	-5,797

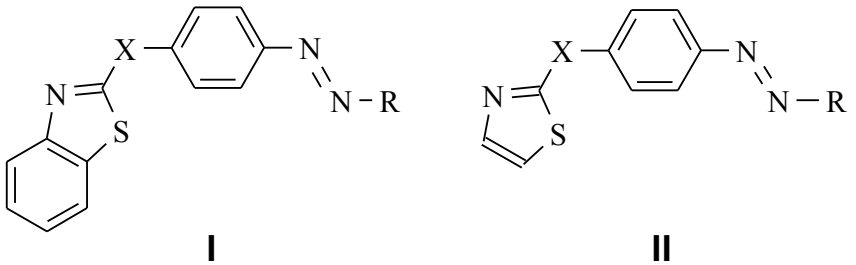
[a] *5-α*
[b] *5-β*
[c] *Na węglu C_{10} zamiast H jest grupa Me*

5.4.1.2 Barwniki heterocykliczne o powinowactwie do celulozy

Oddziaływanie barwników z celulozą ma odmienny charakter niż oddziaływanie leków z receptorem. Do opisu tego skomplikowanego zjawiska używana jest izoterma Langmuira [93, 94]. Takie podejście nie wyjaśnia jednak molekularnych podstaw oddziaływania.

Badania nad oddziaływaniem barwników z celulozą skłaniają do założenia, że barwniki wiążą się z celulozą w ściśle uporządkowany sposób [95]. Opublikowano kilka prac poświęconych modelowaniu QSAR powinowactwa barwników do celulozy [32, 96, 97, 98, 99, 100, 101, 102, 103, 104]. Wydaje się więc, że zastosowanie farmakoforowych koncepcji modelowania jest uzasadnione. W takim podejściu celuloza pełni rolę swoistego receptora. Jest to jednak szczególny receptor – miejsce wiązania się barwników nie jest w celulozie ograniczone do pojedynczej kieszeni receptorowej bądź miejsca aktywnego. Cząsteczka celulozy jest naturalnym polimerem posiadającym wiele powtarzających się miejsc mogących oddziaływać z cząsteczkami barwnika. Oddziaływanie poszczególnych cząsteczek jest jednak najprawdopodobniej specyficzne [105].

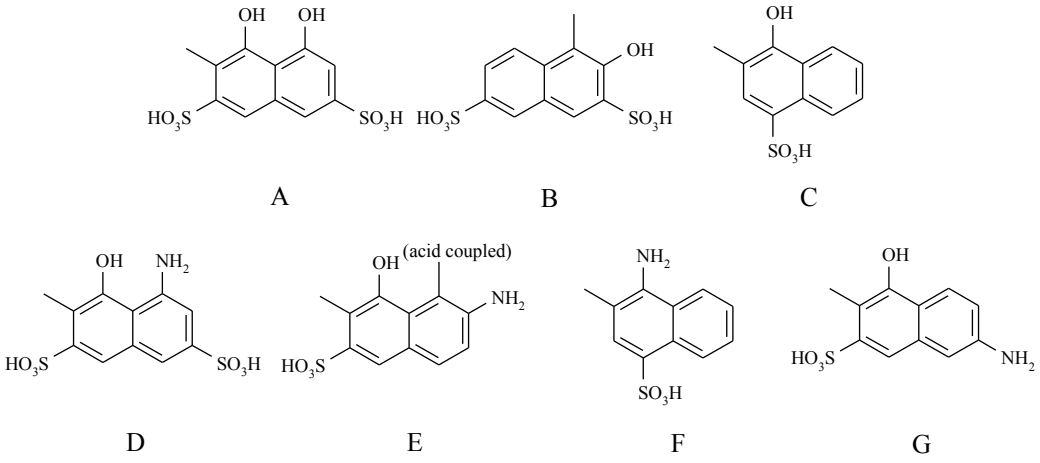
Metoda s-CoMSA została użyta do modelowania grupy 21 barwników heterocyklicznych [32, 106]. Wszystkie cząsteczki szeregu zostały poddane optymalizacji geometrii. Struktury związków znajdują się w tabeli 5.2 oraz na rysunku 5.4 oraz 5.5. Przetestowano trzy różne sposoby (a, b, c) superpozycji molekuł. Szczegóły przedstawiono na rysunku 5.6. Maksymalna wartość parametru q^2_{CV} uzyskana dla modelu obejmującego wszystkie cząsteczki szeregu wyniosła 0,97 i jest porównywalna z wartością uzyskaną dla modelu SOM-CoMSA (q^2_{CV} = 0,98) [32].



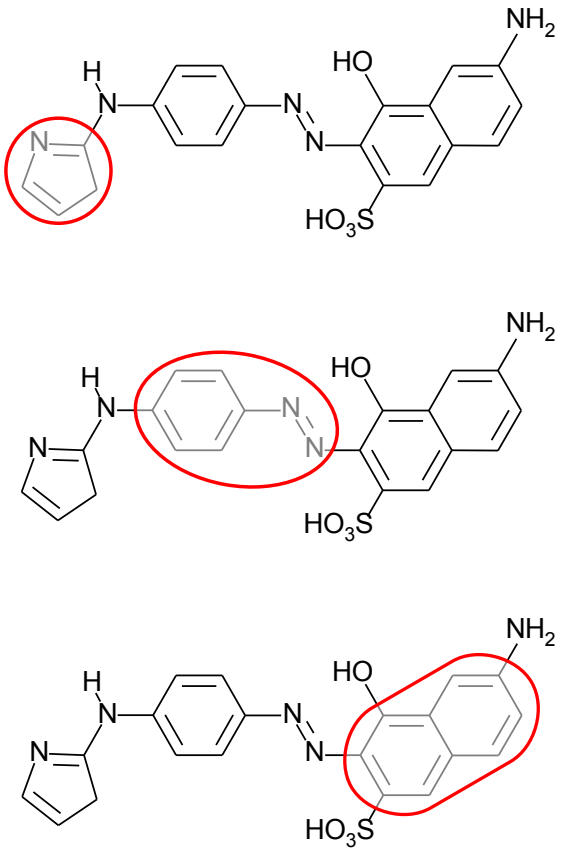
I **II**

Rysunek 5.4 Szkielety analizowanych heterocyklicznych barwników azowych. Podpisy odpowiadają kolumnie **G** tabeli 5.2, grupy X oraz R odpowiadają kolejno kolumnom **X** oraz **R** tej samej tabeli. Budowa grup R jest przedstawiona na rysunku 5.5.

Modelowanie s-CoMSA było wykonane także z pominięciem związku **d21** oraz po podziale na zbiór modelowy i testowy. Zbiór modelowy zawierał związki nieparzyste **d1**, **d3**, **d5**, ..., **d21**, a zbiór testowy związki parzyste: **d2**, **d4**, **d6**, ..., **d20**. Szczegółowe omówienie uzyskanych wyników znajduje się w rozdziałach 5.4.2 (strona 60), 5.4.3 (strona 61) oraz 7.1.1 (strona 83).



Rysunek 5.5 Rodzaje grup R (patrz rysunek 5.4). Podpisy odpowiadają kolumnie **R** tabeli 5.2.



Rysunek 5.6 Wzorzec nakładania i sposób przeprowadzania superpozycji szeregu barwników azowych. Czerwonym obrysem zaznaczono atomy, na które nakładane były analogiczne cząsteczki szeregu. Każdy z zaznaczonych motywów występuje we wszystkich cząsteczkach szeregu.

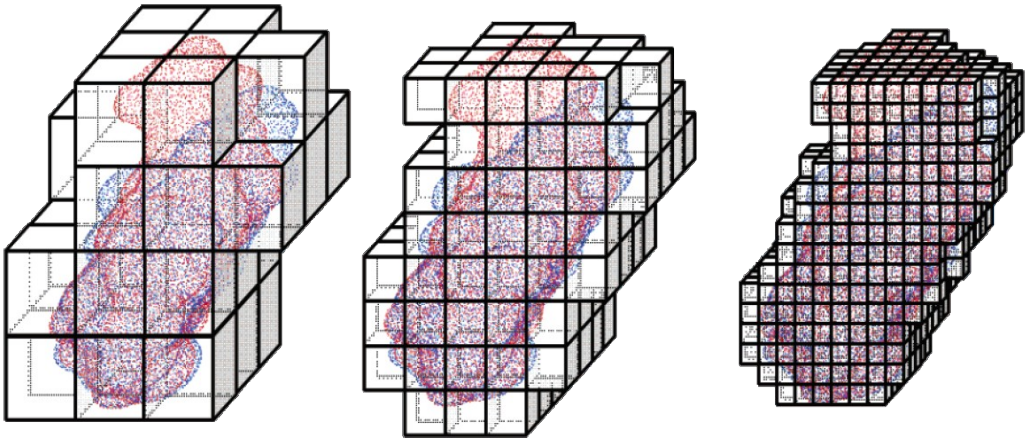
Tabela 5.2 Struktury analizowanych barwników azowych oraz ich powinowactwo do włókna. Kolumna **G** oznacza jeden z dwóch szkieletów zamieszczonych na rysunku 5.4. Kolumny **X** oraz **R** odpowiadają odpowiednimi grupom funkcyjnym zaznaczonym na tym samym rysunku. Oznaczenia grup kolumny **R** znajdują się na rysunku 5.5. Ostatnia kolumna zawiera powinowactwo do włókna.

Nr	G	X	R	$-\Delta\mu^0$ (kJ/mol)
d1	I	-NH-	A	6,78
d2	I	-NH-	B	9,20
d3	I	-NH-	D	12,60
d4	I	-NH-	E	15,30
d5	I	O	A	3,26
d6	I	O	B	5,27
d7	I	O	D	7,61
d8	I	O	E	10,30
d9	I	O	G	10,20
d10	I	S	A	1,26
d11	I	S	B	3,56
d12	I	S	D	5,02
d13	I	S	E	8,45
d14	I	S	G	8,12
d15	II	-NH-	E	15,33
d16	II	-NH-	D	12,60
d17	II	-NH-	B	9,24
d18	II	-NH-	A	6,80
d19	I	S	C	5,86
d20	I	S	F	10,33
d21	I	S	E*	9,75

* grupa połączona jest przez atom 5

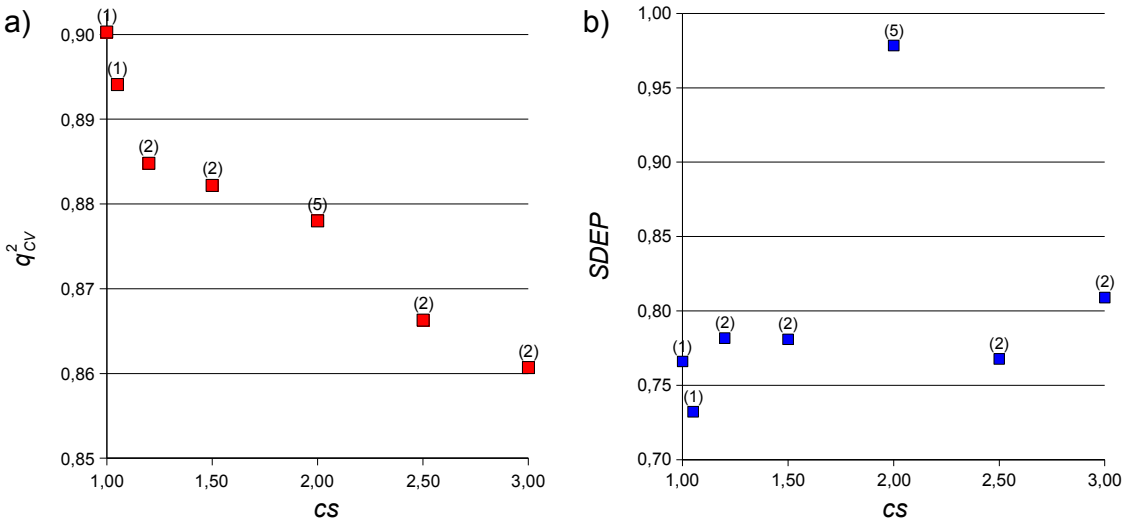
5.4.2 Wpływ rozdzielczości siatki na modelowanie s-CoMSA

Podstawowym parametrem charakteryzującym deskryptor s-CoMSA jest rozmiar pojedynczej komórki siatki, tzw. parametr *cs*. Parametr *cs* określa rozdzielczość reprezentacji powierzchni w postaci deskryptora s-CoMSA. Wraz ze zmniejszaniem parametru *cs* dokładność odwzorowania trójwymiarowej struktury rośnie. Zwiększa się również liczba sektorów koniecznych do opisania struktury związku. Rysunek 5.7 przedstawia tę zależność. Wzrost liczby sektorów zależy w trzeciej potędze od zmiany wielkości *cs*. Dwukrotne zmniejszenie *cs* powoduje ośmiokrotny wzrost liczby sektorów.



Rysunek 5.7 Dokładność charakterystyki powierzchni jest większa dla małych sektorów. Wraz ze zmniejszaniem sektorów wzrasta jednak ich liczba.

Na rysunku 5.8 przedstawiono wpływ rozdzielczości siatki na efektywność modelowania aktywności szeregu steroidów CBG. Wielkość sektora zmieniała się w granicach od 1 do 3 Å. Wraz ze wzrostem rozmiaru sektora q^2_{CV} nieznacznie maleje. Zmiana wielkości sektora od 1 do 3 Å powoduje zmianę q^2_{CV} o około 0,04 jest więc praktycznie bez znaczenia. Zależność parametru *SDEP* od *cs* potwierdza te spostrzeżenia. W badanym przedziale wartość *SDEP* waha się w nieznacznym zakresie. Jedynie dla *cs* = 2 Å występuje niezgodność. Jest ona jednak spowodowana dużą różnicą między kompleksowością tego modelu a kompleksowością pozostałych modeli.

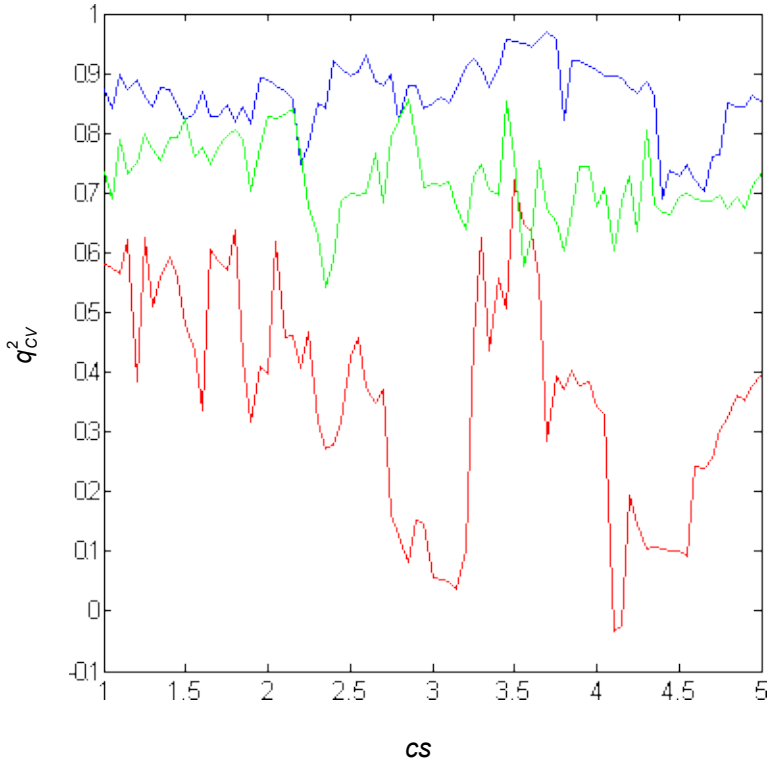


Rysunek 5.8 Zależność parametrów q^2_{CV} (a) oraz *SDEP* (b) od rozmiaru sektora *cs* dla grupy steroidów CBG **s1** do **s21** modelowanych metodą s-CoMSA. W nawiasach podano kompleksowość modeli.

Zdolność prognozowania modeli uzyskanych dla sektora 1 Å jest praktycznie równa zdolności modeli uzyskanych dla siatki 1,5 Å a liczba sektorów koniecznych do charakterystyki powierzchni przy takiej różnicy rozmiaru wzrasta ponad 3 razy. Uzyskano modele o dobrej zdolności prognozowania stosując również sektory o wielkości rzędu 3-4 Å. Mimo tego standardowy rozmiar sektora został ustalony na 1 Å, co umożliwia lepszą wizualizację modeli s-CoMSA. Mniejsze sektory umożliwiają precyzyjniejsze wskazywanie różnych obszarów na powierzchni cząsteczek zidentyfikowanych w toku dalszej analizy s-CoMSA.

Dokładne badania zależności q^2_{CV} od rozmiaru sektora przeprowadzono również dla grupy heterocyklicznych barwników azowych (zobacz rozdział 5.4.1.2). Wykonano szereg modeli w oparciu o cały zbiór barwników stosując różną gęstość siatki. Rozmiar sektora wahał się w granicach od 1 do 5 Å, krok wynosił 0,05 Å. Wykres zależności q^2_{CV} od wielkości sektora dla wszystkich trybów nakładania znajduje się na rysunku 5.9.

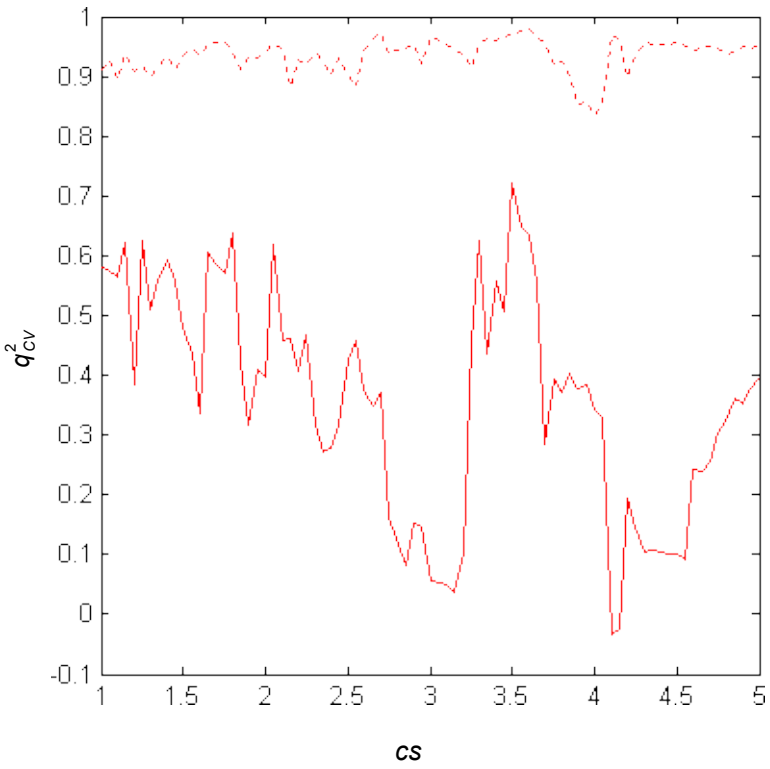
Wartość parametru q^2_{CV} dla superpozycji a oraz b ulega niewielkim wahaniom w badanym przedziale wartości parametru *cs*. Natomiast w przypadku trybu c wartość q^2_{CV} jest niestabilna – wykazuje silne wahanie w zależności od rozmiaru sektora. Dodatkowo poziom wartości q^2_{CV} dla tego trybu jest wyraźnie niższy niż w przypadku pozostałych trybów.



Rysunek 5.9 Zależność parametru q^2_{CV} od rozmiaru sektora *cs* oraz trybu nakładania szeregu heterocyklicznych barwników azowych. Kolor zielony odpowiada procedurze nakładania a, kolor niebieski trybowi b, kolor czerwony trybowi c.

5.4.3 Wpływ superpozycji cząsteczek

Istotnym etapem analizy s-CoMSA jest superpozycja trójwymiarowych struktur. Pogorszenie zdolności prognozowania modeli QSAR barwników heterocyklicznych, uzyskanych dla trybu nakładania c, jest spowodowane nieprawidłowym nałożeniem związku **d21** (patrz rysunek 5.9, strona 61).



Rysunek 5.10 Zależność parametru q^2_{cv} od rozmiaru sektora *cs* dla procedury nakładania c całego szeregu heterocyklicznych barwników azowych (linia ciągła) oraz dla szeregu bez związku **d21** (linia przerywana). Po odrzuceniu związku **d21** widoczna jest znaczna poprawa ogólnego poziomu oraz stabilności parametru q^2_{cv}.

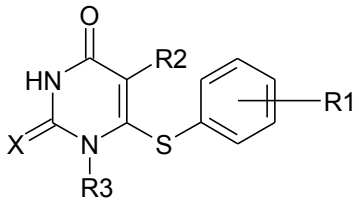
Wyeliminowanie ze zbioru związku **d21** i powtórzenie analizy dla procedury nakładania c wyraźnie poprawia stabilność oraz ogólny poziom wartości q^2_{CV} (rysunek 5.10). Związek **d21** ma odmienną strukturę od pozostałych cząsteczek szeregu. Grupa funkcyjna R (patrz tabela 5.2 oraz rysunki 5.4, 5.5) jest połączona w sposób odmienny od reszty związków. Odmienność związku **d21** jest szczególnie wyraźna w trybie nakładania c, w którym cząsteczki są nakładane właśnie na atomy grupy funkcyjnej R (patrz rysunek 5.6).

Wpływ trybu superpozycji na wyniki modelowania s-CoMSA badany był również dla inhibitorów reduktazy kwasu foliowego, pochodnych 2,4-diamino-5-benzylpirymidyny (patrz rozdział 6.2, strona 68). Testowano trzy tryby nakładania a, b i c. Modele uzyskane dla poszczególnych trybów nie różniły się znacznie pod względem zdolności prognozowania.

6 Aplikacje metody s-CoMSA

6.1 Modelowanie aktywności hamowania odwrotnej transkryptazy HIV pochodnych 1[2-(hydroksyetoksy)metylo]-6(fenylotio)tyminy – HEPT

Pochodne HEPT (ang. hydroxyethoxy phenylthio thymine) należą do nienukleozydowych inhibitorów odwrotnej transkryptazy wirusa HIV, dla których opisano wiele modeli QSAR [107, 108, 109, 110, 111, 112, 113, 114]. W tabeli 6.1 przedstawiono budowę 107 reprezentatywnych związków wybranych spośród modeli opisanych w literaturze [107].



Rysunek 6.1 Szkielet 1[2-(hydroksyetoksy)metylo]-6(fenylotio)tyminy. Symbole grup funkcyjnych X, R1, R2, R3 odpowiadają kolumnom **X**, **R1**, **R2**, **R3** tabeli 6.1.

Tabela 6.1 Struktury i aktywność analizowanych pochodnych HEPT. Kolumny **X**, **R1**, **R2**, **R3** odpowiadają grupom funkcyjnym na rysunku 6.1. Ostatnia kolumna podaje aktywność hamowania odwrotnej transkryptazy HIV.

Nr	R1	R2	R3	X	Aktywność [log 1/IC_{50}]
h1	2-Me	Me	$CH_2OCH_2CH_2OH$	O	4,15
h2	2-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3,85
h3	2-OMe	Me	$CH_2OCH_2CH_2OH$	O	4,72
h4	3-Me	Me	$CH_2OCH_2CH_2OH$	O	5,59
h5	3-Et	Me	$CH_2OCH_2CH_2OH$	O	5,57
h6	3-t-Bu	Me	$CH_2OCH_2CH_2OH$	O	4,92
h7	3-CF_3	Me	$CH_2OCH_2CH_2OH$	O	4,35
h8	3-F	Me	$CH_2OCH_2CH_2OH$	O	5,48
h9	3-Cl	Me	$CH_2OCH_2CH_2OH$	O	4,89
h10	3-Br	Me	$CH_2OCH_2CH_2OH$	O	5,24
h11	3-I	Me	$CH_2OCH_2CH_2OH$	O	5,00

Nr	R1	R2	R3	X	Aktywność [log 1/IC_{50}]
h12	3-NO_2	Me	$CH_2OCH_2CH_2OH$	O	4,47
h13	3-OH	Me	$CH_2OCH_2CH_2OH$	O	4,09
h14	3-OMe	Me	$CH_2OCH_2CH_2OH$	O	4,66
h15	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	O	6,59
h16	3,5-Cl_2	Me	$CH_2OCH_2CH_2OH$	O	5,89
h17	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	S	6,66
h18	3-COOMe	Me	$CH_2OCH_2CH_2OH$	O	5,10
h19	3-COMe	Me	$CH_2OCH_2CH_2OH$	O	5,14
h20	3-CN	Me	$CH_2OCH_2CH_2OH$	O	5,00
h21	H	$CH_2CH=CH_2$	$CH_2OCH_2CH_2OH$	O	5,60
h22	H	Et	$CH_2OCH_2CH_2OH$	S	6,96
h23	H	Pr	$CH_2OCH_2CH_2OH$	S	5,00
h24	H	i-Pr	$CH_2OCH_2CH_2OH$	S	7,23
h25	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	S	8,11
h26	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	S	8,30
h27	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	S	7,37
h28	H	Et	$CH_2OCH_2CH_2OH$	O	6,92
h29	H	Pr	$CH_2OCH_2CH_2OH$	O	5,47
h30	H	i-Pr	$CH_2OCH_2CH_2OH$	O	7,20
h31	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	O	7,89
h32	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	O	8,57
h33	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	O	7,85
h34	4-Me	Me	$CH_2OCH_2CH_2OH$	O	3,66
h35	H	Me	$CH_2OCH_2CH_2OH$	O	5,15
h36	H	Me	$CH_2OCH_2CH_2OH$	S	6,01
h37	H	I	$CH_2OCH_2CH_2OH$	O	5,44
h38	H	$CH=CH_2$	$CH_2OCH_2CH_2OH$	O	5,69
h39	H	CH=CHPh	$CH_2OCH_2CH_2OH$	O	5,22
h40	H	CH_2Ph	$CH_2OCH_2CH_2OH$	O	4,37
h41	H	$CH=CPh_2$	$CH_2OCH_2CH_2OH$	O	6,07
h42	H	Me	$CH_2OCH_2CH_2OMe$	O	5,06
h43	H	Me	$CH_2OCH_2CH_2OAc$	O	5,17
h44	H	Me	$CH_2OCH_2CH_2OCOPh$	O	5,12
h45	H	Me	CH_2OCH_2Me	O	6,48

Nr	R1	R2	R3	X	Aktywność [log 1/IC_{50}]
h46	H	Me	$CH_2OCH_2CH_2Cl$	O	5,82
h47	H	Me	$CH_2OCH_2CH_2N_3$	O	5,24
h48	H	Me	$CH_2OCH_2CH_2F$	O	5,96
h49	H	Me	$CH_2OCH_2CH_2Me$	O	5,48
h50	H	Me	CH_2OCH_2Ph	O	7,06
h51	H	Et	CH_2OCH_2Me	O	7,72
h52	H	Et	CH_2OCH_2Me	S	7,58
h53	3,5-Me_2	Et	CH_2OCH_2Me	O	8,24
h54	3,5-Me_2	Et	CH_2OCH_2Me	S	8,30
h55	H	Et	CH_2OCH_2Ph	O	8,23
h56	3,5-Me_2	Et	CH_2OCH_2Ph	O	8,55
h57	H	Et	CH_2OCH_2Ph	S	8,09
h58	3,5-Me_2	Et	CH_2OCH_2Ph	S	8,14
h59	H	i-Pr	CH_2OCH_2Me	O	7,99
h60	H	i-Pr	CH_2OCH_2Ph	O	8,51
h61	H	i-Pr	CH_2OCH_2Me	S	7,89
h62	H	i-Pr	CH_2OCH_2Ph	S	8,14
h63	H	Me	CH_2OMe	O	5,68
h64	H	Me	CH_2OBu	O	5,33
h65	H	Me	Et	O	5,66
h66	H	Me	Bu	O	5,92
h67	3,5-Cl_2	Et	CH_2OCH_2Me	S	7,89
h68	H	Et	CH_2O-i-Pr	S	6,66
h69	H	Et	CH_2O-c-Hex	S	5,79
h70	H	Et	CH_2OCH_z-c-Hex	S	6,45
h71	H	Et	$CH_2OCH_2C_6H_4$(4-Me)	S	7,11
h72	H	Et	$CH_2OCH_2C_6H_4$(4-Cl)	S	7,92
h73	H	Et	$CH_2OCH_2CH_2Ph$	S	7,04
h74	3,5-Cl_2	Et	CH_2OCH_2Me	O	8,13
h75	H	Et	CH_2O-i-Pr	O	6,47
h76	H	Et	CH_2O-c-Hex	O	5,40
h77	H	Et	CH_2OCH_z-c-Hex	O	6,35
h78	H	Et	$CH_2OCH_zCH_2Ph$	O	7,02
h79	H	c-Pr	CH_2OCH_2Me	S	7,02

Nr	R1	R2	R3	X	Aktywność [log 1/IC_{50}]
h80	H	c-Pr	CH_2OCH_2Me	O	7,00
h81	H	Me	$CH_2OCH_2CH_2OC_5H_{11}$-n	O	4,46
h82	2-Cl	Me	$CH_2OCH_2CH_2OH$	O	3,89
h83	3-CH_2OH	Me	$CH_2OCH_2CH_2OH$	O	3,53
h84	4-F	Me	$CH_2OCH_2CH_2OH$	O	3,60
h85	4-Cl	Me	$CH_2OCH_2CH_2OH$	O	3,60
h86	4-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3,72
h87	4-CN	Me	$CH_2OCH_2CH_2OH$	O	3,60
h88	4-OH	Me	$CH_2OCH_2CH_2OH$	O	3,56
h89	4-OMe	Me	$CH_2OCH_2CH_2OH$	O	3,60
h90	4-COMe	Me	$CH_2OCH_2CH_2OH$	O	3,96
h91	4-COOH	Me	$CH_2OCH_2CH_2OH$	O	3,45
h92	3-$CONH_2$	Me	$CH_2OCH_2CH_2OH$	O	3,51
h93	H	COOMe	$CH_2OCH_2CH_2OH$	O	5,18
h94	H	CONHPh	$CH_2OCH_2CH_2OH$	O	4,74
h95	H	SPh	$CH_2OCH_2CH_2OH$	O	4,68
h96	H	CCH	$CH_2OCH_2CH_2OH$	O	4,74
h97	H	CCPh	$CH_2OCH_2CH_2OH$	O	5,47
h98	3-NH_2	Me	$CH_2OCH_2CH_2OH$	O	3,60
h99	H	$COCHMe_2$	$CH_2OCH_2CH_2OH$	O	4,92
h100	H	COPh	$CH_2OCH_2CH_2OH$	O	4,89
h101	H	CCMe	$CH_2OCH_2CH_2OH$	O	4,72
h102	H	F	$CH_2OCH_2CH_2OH$	O	4,00
h103	H	Cl	$CH_2OCH_2CH_2OH$	O	4,52
h104	H	Br	$CH_2OCH_2CH_2OH$	O	4,70
h105	H	Me	$CH_2OCH_2CH_2OCH_2Ph$	O	4,70
h106	H	Me	H	O	3,60
h107	H	Me	Me	O	3,82

Optymalny model obejmujący wszystkie związki szeregu charakteryzował się q^2_{CV} równym 0,86. Został on uzyskany dla siatki o gęstości 1,5 Å po zastosowaniu procedury eliminacji zmiennych IVE-PLS. W celu walidacji zdolności prognozowania szereg pochodnych HEPT został podzielony podobnie jak w innych pracach na zbiory modelowy i testowy zawierające odpowiednio związki **h1**-**h80** oraz **h81**-**h107** [107, 110]. Dodatkowo w celu znalezienia reprezentatywnego podziału na zbiór modelowy i testowy zastosowano metodę Kennarda-Stone'a [115]. Porównanie wyników modelowania metodą s-CoMSA z wynikami modelowania innymi metodami QSAR znajduje się w tabeli 6.2. Omówienie wyników walidacji znajduje się w rozdziale 8.1 (patrz strona 104).

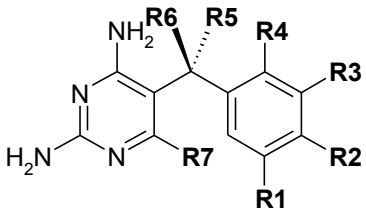
Tabela 6.2 Wyniki modelowania aktywności pochodnych HEPT różnymi metodami QSAR.

Model	q^2_{CV}	***SDEP***
s-CoMSA-107-IVE [a]	0,86	---
s-CoMSA-80/27 [a, d]	0,79	2,01
s-CoMSA-KS [a, e]	0,72	0,69
4D-QSAR-J_q [b]	0,77	1,46
4D-QSAR-J_q-IVE [b]	0,95	1,48
SOM-4D-QSAR$_q$ [b]	0,77	1,76
SOM-4D-QSAR$_q$-IVE [b]	0,98	1,68
CoMFA-101 [c]	0,86	---

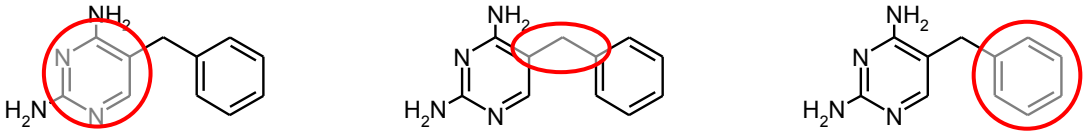
a) gęstość siatki 1,5 Å
b) dane zgodnie z publikacją [87]
c) dane zgodnie z publikacją [109]
d) podział na zbiór modelowy **h1**-**h80** i testowy **h81**-**h107**
e) podział na zbiór modelowy i testowy w proporcjach 80/27 metodą Kennard-Stone patrz rozdział 8.1, strona 104

6.2 Modelowanie aktywności inhibitorów reduktazy kwasu foliowego – pochodnych 2,4-diamino-5-benzylpirymidyny

Kwas foliowy odgrywa istotną rolę w procesach związanych z powstawaniem nowych komórek *in vivo*. Reduktaza kwasu foliowego przekształca go w aktywną formę, niezbędną do rozwoju komórek. Efekt terapeutyczny inhibitorów tego enzymu polega na hamowaniu wzrostu komórek nowotworowych związanego z niedoborem aktywniej formy kwasu foliowego. W literaturze opisano wiele modeli QSAR związanych z opisanymi efektami w tym także modele 4D-QSAR [30, 48]. Struktury modelowanych inhibitorów oraz ich aktywności przedstawiono w tabeli 6.3. Wygenerowane trójwymiarowe struktury zostały nałożone na siebie na trzy różne sposoby. Wzorzec nakładania oraz sposób nałożenia pokazano na rysunku 6.3.



Rysunek 6.2 Analizowane inhibitory reduktazy kwasu foliowego. Symbole grup funkcyjnych od R1 do R7 odpowiadają kolumnom tabeli 6.3 od **R1** do **R7**.



Rysunek 6.3 Superpozycja a, b, c. Kolorem czerwonym zaznaczono atomy, na które nakładane były cząsteczki analizowanego szeregu.

Tabela 6.3 Struktury analizowanych inhibitorów reduktazy kwasu foliowego. Kolumny od R1 do R7 oznaczają różne pozycje grup funkcyjnych. Stosowne objaśniania znajdują się na rysunku 6.2. Ostatnia kolumna zawiera wartości hamowania enzymu wyrażone jako logarytm odwrotności stężenia wywołującego 50% hamowanie enzymu.

Nr	R1	R2	R3	R4	R5	R6	R7	$\log(1/I_{50})$
f1	OCH_3	OCH_3	OCH_3	H	H	H	H	8,23
f2	OCH_3	OCH_3	OCH_3	CH_3	H	H	H	5,85
f3*R*	OCH_3	OCH_3	OCH_3	H	OH	CH_3	H	4,00
f4*S*	OCH_3	OCH_3	OCH_3	H	OH	CH_3	H	4,00
f5	OCH_3	OCH_3	OCH_3	H	$=CH_2$		H	5,60
f6*R*	OCH_3	OCH_3	OCH_3	H	H	CH_3	H	5,35
f7*S*	OCH_3	OCH_3	OCH_3	H	H	CH_3	H	5,35
f8	OCH_3	Br	OCH_3	H	H	H	H	8,53
f9	OCH_3	OH	OCH_3	H	H	H	H	7,96
f10	OCH_3	OH	OCH_3	H	H	H	CH_3	6,52
f11	OCH_3	OCH_3	OCH_3	H	H	H	CH_3	7,00
f12	OH	H	OH	H	H	H	H	2,78
f13	H	H	H	H	H	H	H	5,71
f14	CH_2OH	H	CH_3OH	H	H	H	H	5,83
f15	H	H	Cl	H	H	H	H	6,14
f16	H	Br	H	H	H	H	H	6,30
f17	OCH_3	H	H	H	H	H	H	6,40
f18	OCH_3	H	OCH_3	H	H	H	H	7,75
f19	CH_3	H	CH_3	H	H	H	H	7,45
f20	H	C_6H_5	H	H	H	H	H	6,40

Szereg podzielono na dwa zbiory: modelowy i testowy. Zbiór modelowy zawierał związki od **f1** do **f11** oraz od **f13** do **f15**. W zbiorze testowym znalazły się związki od **f16** do **f20**. Związek **f12** został wykluczony z analizy ponieważ jest on obiektem odległym. Jako jedyny posiada grupy OH w pozycjach R1 i R3. Wykazuje również najniższą aktywność. Uzyskane modele s-CoMSA przedstawiono w tabeli 6.4. Porównano je z analogicznymi modelami SOM-CoMSA.

Tabela 6.4 Wyniki modelowania QSAR zbioru inhibitorów reduktazy kwasu foliowego – pochodnych 2,4-diamino-5-benzylpirymidyny.

Podział na zbiory	Parametry	Nakładanie		
		a	b	c
s-CoMSA				
Zbiór modelowy i treningowy	*cs* *	1	1	1
	q^2_{CV}	0,54	0,70	0,69
	s_{CV}	1,96	1,59	1,26
	SDEP	1,42	1,32	1,42
Zbiór modelowy i treningowy IVE	*cs* *	1	1	4
	q^2_{CV}	0,73	0,59	0,68
	s_{CV}	0,83	0,92	0,85
	SDEP	0,93	1,15	1,10
Cały zbiór	*cs* *	1	1	3
	q^2_{CV}	0,38	0,56	0,47
	s_{CV}	1,50	0,90	1,01
SOM-CoMSA				
Zbiór modelowy i treningowy	q^2_{CV}	0,59	0,64	0,71
	s_{CV}	1,02	0,91	0,90
	SDEP	1,25	0,96	1,20
Zbiór modelowy i treningowy IVE	q^2_{CV}	0,62	0,87	0,71
	s_{CV}	0,89	0,58	0,79
	SDEP	0,81	0,72	0,78
Cały zbiór	q^2_{CV}	0,62	0,72	0,64
	s_{CV}	1,17	1,01	1,08

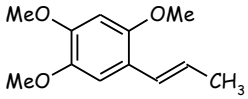
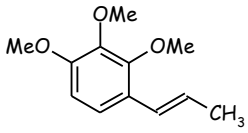
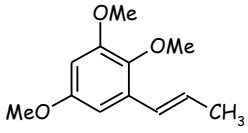
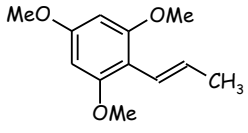
* Rozmiar sektora – parametr *cs*

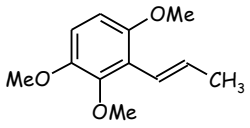
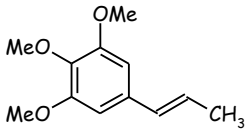
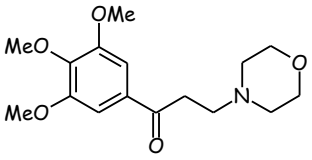
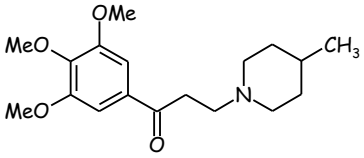
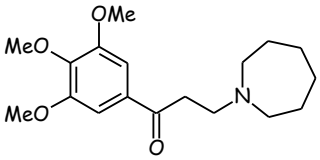
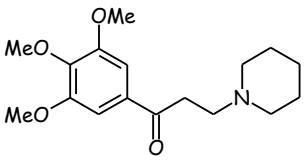
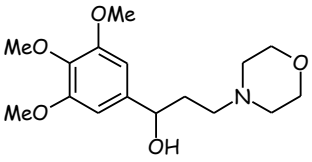
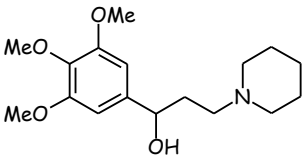
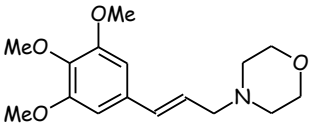
6.3 Modelowanie aktywności pochodnych α-asaronu

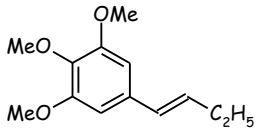
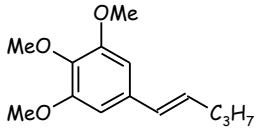
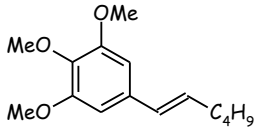
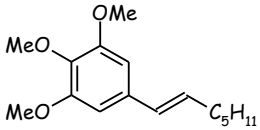
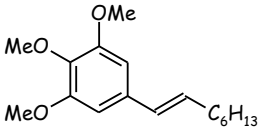
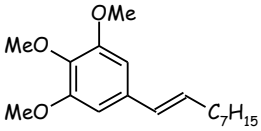
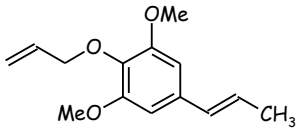
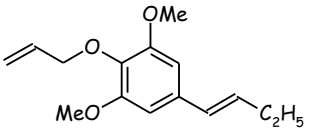
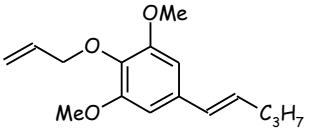
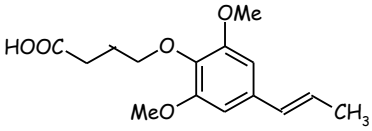
Miażdżyca i towarzyszące jej schorzenia są jednym z poważniejszych zagorzeń życia w rejonach wysoko uprzemysłowionych. Istnieje wiele czynników wywołujących tego typu schorzenia. Zależność między chorobami układu krążenia a wysokim poziomem cholesterolu jest oczywista [116]. Obniżanie poziomu lipoprotein VLDL-C (ang. very low density lipoprotein) oraz LDL-C (ang. low density lipoprotein) jest podstawowym celem terapii [117]. Coraz bardziej oczywista staje się również konieczność regulacji poziomów innych lipoprotein takich jak HDL-C (ang. high density lipoprotein) oraz triglicerydów (TG) [118].

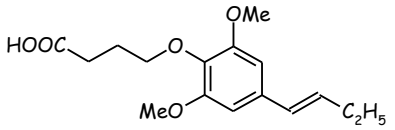
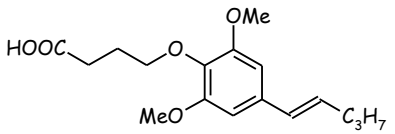
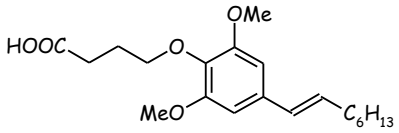
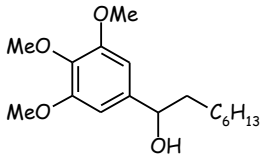
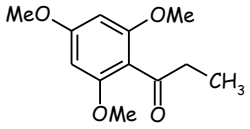
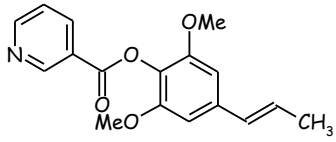
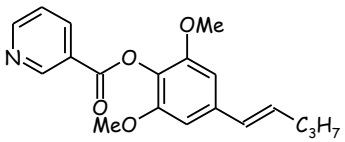
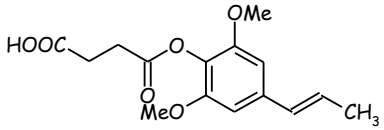
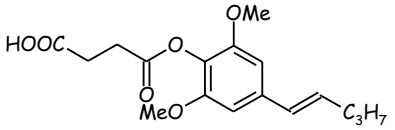
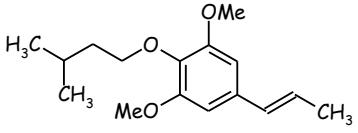
Pochodne α-asaronu wykazują działanie hipolipidemiczne [119, 120]. Obniżają one poziom stężenia niepożądanych frakcji lipoprotein bez istotnej zmiany stężenia samego cholesterolu. Do modelowania s-CoMSA wybrano 40 pochodnych α-asaronu [28]. Struktury związków znajdują się w tabeli 6.5. Miarą aktywności pochodnych α-asaronu jest indeks aterogenny $I_{TG/HDL}$. Uzyskuje się go przez podzielenie wywoływanych poziomów stężeń triglicerydów TG oraz lipoprotein HDL-C.

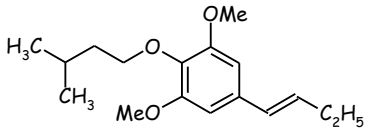
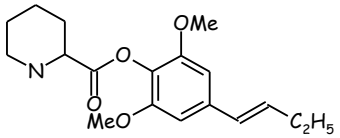
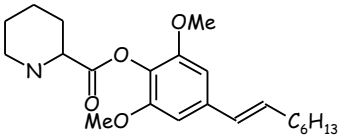
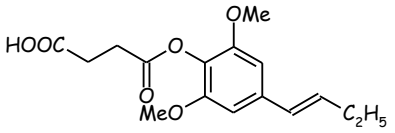
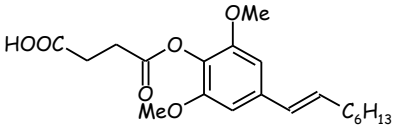
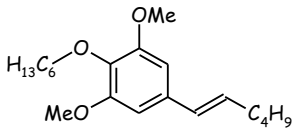
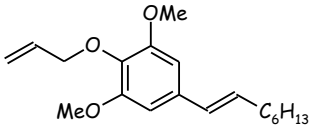
Tabela 6.5 Struktury analizowanych pochodnych α-asaronu. Indeks aterogenny obliczony na podstawie kolumn **TG** oraz **HDL-C**.

Nr	Struktura	$I_{TG/HDL}$	TG [mmol/L]	HDL-C [mmol/L]
a1		2,10	2,48	1,18
a2		1,27	1,41	1,11
a3		1,56	1,58	1,01
a4		1,34	1,31	0,98

Nr	Struktura	$I_{TG/HDL}$	TG [mmol/L]	HDL-C [mmol/L]
a5		2,35	1,83	0,78
a6		1,31	1,53	1,17
a7		2,00	1,06	0,53
a8		1,38	0,58	0,42
a9		1,48	0,77	0,52
a10		1,71	0,87	0,51
a11		1,26	0,53	0,42
a12		1,74	0,59	0,34
a13		2,05	0,8	0,39

Nr	Struktura	$I_{TG/HDL}$	TG [mmol/L]	HDL-C [mmol/L]
a14		0,36	0,21	0,59
a15		0,45	0,24	0,53
a16		0,16	0,09	0,57
a17		0,27	0,19	0,71
a18		0,13	0,07	0,56
a19		0,14	0,08	0,57
a20		0,47	0,2	0,43
a21		0,15	0,1	0,66
a22		0,10	0,07	0,73
a23		0,45	0,29	0,64

Nr	Struktura	$I_{TG/HDL}$	TG [mmol/L]	HDL-C [mmol/L]
a24		0,22	0,14	0,64
a25		0,13	0,09	0,69
a26		0,85	0,4	0,47
a27		0,77	0,33	0,43
a28		1,90	1,71	0,90
a29		3,15	1,04	0,33
a30		1,52	0,67	0,44
a31		1,79	0,61	0,34
a32		3,28	0,95	0,29
a33		0,45	0,21	0,47

Nr	Struktura	$I_{TG/HDL}$	TG [mmol/L]	HDL-C [mmol/L]
a34		0,56	0,18	0,32
a35		1,98	0,99	0,5
a36		2,58	1,01	0,4
a37		1,92	0,96	0,5
a38		1,96	1,06	0,54
a39		1,28	0,77	0,6
a40		0,80	0,43	0,54

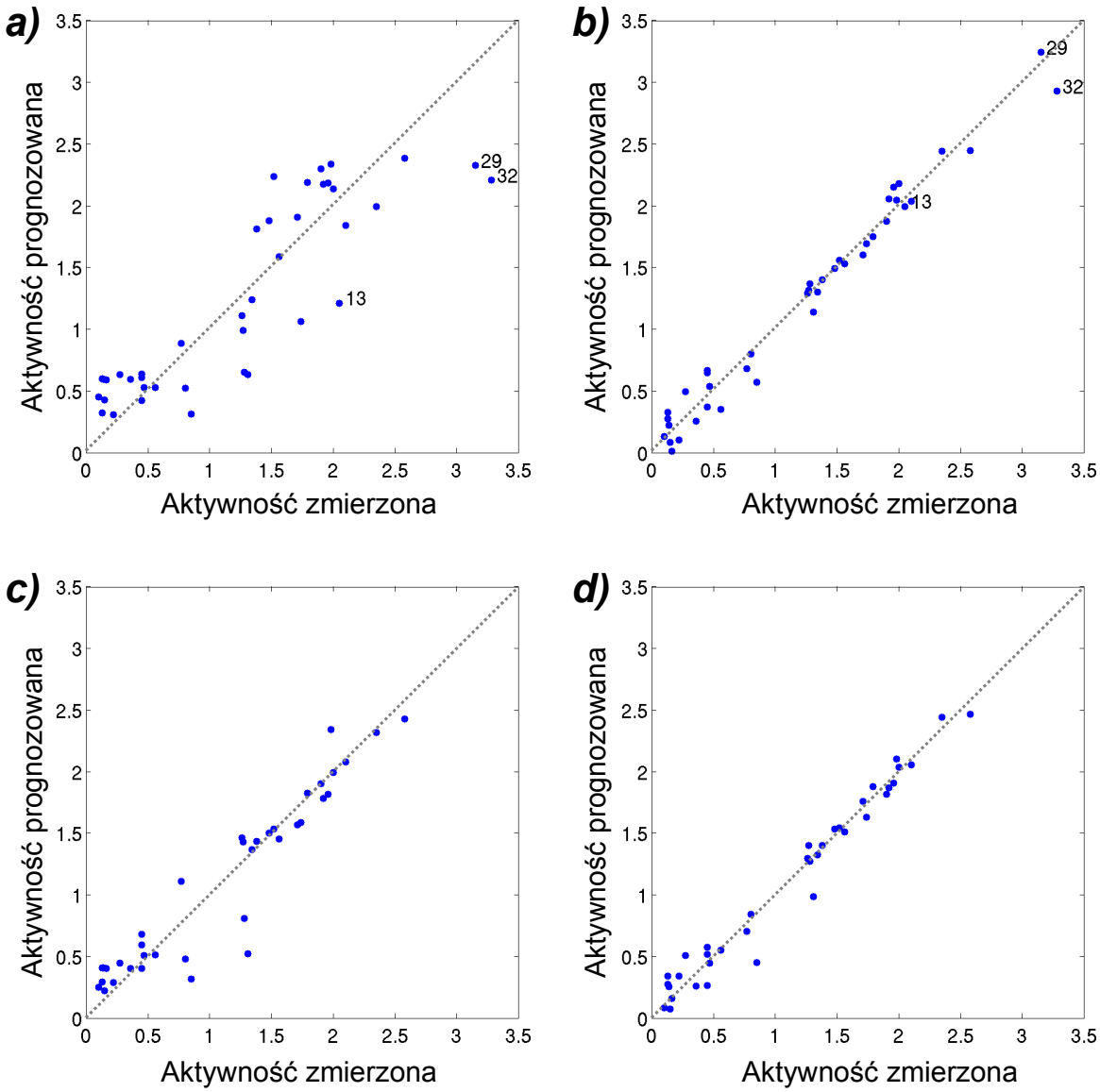
Wyniki modelowania s-CoMSA przedstawiono w tabeli 6.6. Jakość modeli s-CoMSA mierzona parametrem q^2_{CV} jest porównywalna do q^2_{CV} opisującego modele CoMFA i wynosi 0,53 [28]. Wartość ta jest dość niska, tylko nieznacznie przekracza graniczną wartość q^2_{CV} = 0,5, która oddziela modele istotne od nieistotnych.

Odrzucenie trzech pochodnych **a13**, **a29**, **a32** (model 2b) umożliwia znacznie efektywniejsze modelowanie s-CoMSA. W przypadku modeli 3a oraz 3b różnica między wynikami nie jest tak wyraźna. Widoczne jest to również na rysunku 6.4 – wykluczenie pochodnych **a13**, **a29**, **a32** w przypadku modeli uzyskanych po IVE-PLS (część b oraz d rysunku) nie wpływa znacznie na podatność szeregu na modelowanie. Dodatkowo została przetestowana zdolność prognozowania dla zewnętrznego zbioru w oparciu o podział uzyskany metodą Kennard-Stone [115]. Zbiór podzielono na dwa podzbiory liczące po 27 i 13 związków. Pierwszy zbiór posłużył do otrzymania modelu, drugi był wykorzystany wyłącznie do testowania. W tabeli 6.6 znajdują się wartości parametrów q^2_{CV} oraz *SDEP* uzyskane dla użytego podziału, w celach porównawczych zawarto również wyniki uzyskane w analogiczny sposób metodą SOM-CoMSA [28].

Tabela 6.6 Wyniki modelowania QSAR zbioru pochodnych α-asaronu.

Numer modelu	Metoda	q^2_{CV}	*RMS*	*SDEP*
1	CoMFA [a]	0,58	– [e]	– [e]
2a	s-CoMSA	0,53	0,42	– [e]
2b	s-CoMSA [b]	0,69	0,24	– [e]
3a	s-CoMSA-IVE-PLS [f]	0,92	0,13	– [e]
3b	s-CoMSA-IVE-PLS [b, g]	0,94	0,13	– [e]
4	s-CoMSA	0,55	0,19	0,80 [c]
5	SOM-CoMSA	0,60	0,21	0,64 [d]
6a	s-CoMSA [h]	0,45	0,34	0,49
6b	s-CoMSA-IVE-PLS [h]	0,75	0,33	0,30

a) Obliczenia wykonano dla sondy H(+1),
b) Wynik uzyskany po wyeliminowaniu pochodnych **a13**, **a29** i **a32**,
c) Zbiór testowy (metoda K-S): **a1**:**a5**, **a7**, **a11**, **a15**, **a23**, **a28**, **a29**, **a32**, **a36**,
d) Zbiór testowy (metoda K-S): **a14**:**a17**, **a19**, **a22**, **a24**, **a25**, **a29**, **a32**, **a35**, **a37**, **a38**,
e) Nie było obliczane,
f) Uzyskano po wykonaniu IVE-PLS na modelu nr 2a,
g) Uzyskano po wykonaniu IVE-PLS na modelu nr 2b,
h) Zbiór testowy: **a33**:**a40**.

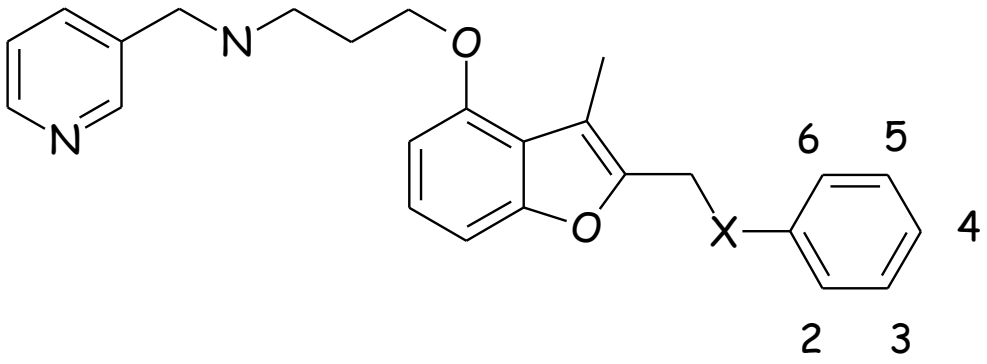


Rysunek 6.4 Zależności między aktywnością zmierzoną a prognozowaną pochodnych α-asaronu. Rysunki a oraz b dotyczą całego zbioru, rysunki c oraz d zbioru bez pochodnych **a13**, **a29**, **a32**. Rysunki a oraz c dotyczą modeli uzyskanych bez wyboru zmiennych – modele 2a oraz 2b; rysunki b oraz d dotyczą modeli uzyskanych po zastosowaniu IVE-PLS – modele 3a oraz 3b (patrz tabela 6.6).

6.4 Modelowanie aktywności benzofuranowych inhibitorów N-mirystytransferazy

N-mirystytransferaza (NMT) jest enzymem katalizującym przemianę kwasu mirystynowego z mirystoilo-CoA do N-terminalnej aminy glicynowej. Jest to proces występujący w wielu organizmach eukariotycznych [121]. Blokowanie enzymu N-mirystytransferazy jest jedną ze strategii zwalczania grzybów *C. Albicans* oraz *C. Neoformans* odpowiedzialnych za różne infekcje i choroby [122, 123]. Interesującą grupą inhibitorów tego enzymu są pochodne benzofuranu. Wykazują one selektywne działanie hamujące wobec N-mirystytransferazy obecnej w grzybach *C. Albicans* [124].

Rysunek 6.5 oraz tabela 6.7 zawierają struktury 29 pochodnych benzofuranu użytych do analizy s-CoMSA [29]. Miarą aktywności był ujemny logarytm parametru IC_{50}. Związki zostały nałożone na siebie względem centralnego pierścienia benzofuranowego. Zastosowanie procedury IVE-PLS pozwoliło uzyskać wysokie wartości q^2_{CV} = 0,96 porównywalne z wartościami uzyskanymi innymi metodami – SOM-CoMSA 0,84, GA-SOM-CoMSA 0,81 [29, 124].



Rysunek 6.5 Wspólny szkielet analizowanych benzofuranowych inhibitorów NMT. Pozycje X, 2, 3, 4, 5, 6 odpow3iadają kolumnom **X**, **R2**, **R3**, **R4**, **R5**, **R6** tabeli 6.7.

Tabela 6.7 Struktury analizowanych benzofuranowych inhibitorów NMT. Kolumny **X**, **R2**, **R3**, **R4**, **R5**, **R6** odpowiadają pozycjom X, 2, 3, 4, 5, 6 zaznaczonym na rysunku 6.5.

Nr	X	R2	R3	R4	R5	R6	log(1/IC$_{50}$)
b1	O	F	H	F	H	H	8,12
b2	O	H	CF_3	H	H	H	6,72
b3	O	H	H	H	H	H	7,14
b4	O	H	H	Cl	H	H	7,14
b5	S	H	H	H	H	H	6,21
b6	S	H	H	Cl	H	H	5,71
b7	O	F	H	H	H	H	8,08
b8	O	H	F	H	H	H	6,95
b9	O	F	H	H	H	F	7,47
b10	O	F	H	H	F	H	8,36
b11	O	F	F	H	H	H	8,44
b12	O	F	F	H	F	H	8,18
b13	O	F	H	F	F	H	8,03
b14	O	F	F	F	H	H	8,24
b15	O	F	F	H	H	F	7,48
b16	O	F	H	F	H	F	7,09
b17	O	F	F	F	F	F	5,85
b18	O	2-Py*					5,53
b19	O	3-Py*					7,24
b20	O	4-Py*					5,81
b21	O	CN	H	H	H	H	7,78
b22	O	H	CN	H	H	H	7,03
b23	S	H	H	F	H	H	5,78
b24	O	F	F	H	F	F	6,24
b25	O	F	H	Br	H	H	7,55
b26	O	H	Br	H	H	H	6,40
b27	O	H	H	Br	H	H	6,06
b28	O	2-Py	Cl	H	H	H	6,49
b29	O	2-Py	H	Cl	H	H	6,64

* Pirydyna

7 Wizualizacja modeli s-CoMSA

7.1 Wybór / eliminacja zmiennych w modelowaniu 3D-QSAR

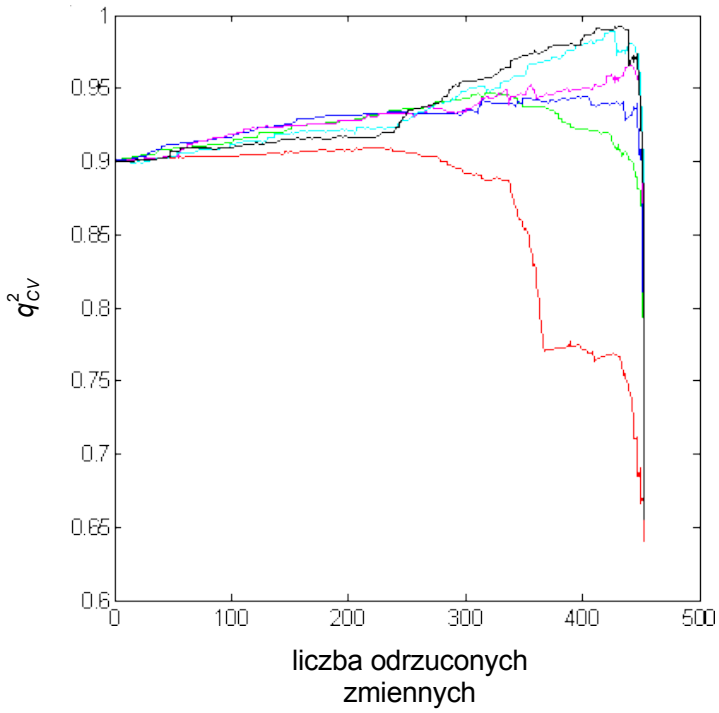
W rozdziale 3.4 omówiono metodę PLS oraz zalety jej stosowania do modelowania wielowymiarowych zależności QSAR. Metoda ta pozwala otrzymać dobrze prognozujące modele. Połączenie PLS z metodami wyboru zmiennych pozwala zwiększyć zdolność prognozowania uzyskiwanych modeli. Najważniejszym celem wyboru zmiennych w badaniach 3D-QSAR jest jednak wskazanie grupy zmiennych związanych z trójwymiarowymi fragmentami struktury, które determinują aktywność analizowanego szeregu. Co więcej, zidentyfikowane fragmenty umożliwiają taką modyfikację struktury, która prowadzi do uzyskania nowej cząsteczki o potencjalnie wyższej aktywności. Uzyskiwanie tego typu informacji jest podstawowym celem metody s-CoMSA oraz innych metod 3D-QSAR [33, 125].

7.1.1 Zastosowanie metody IVE-PLS w modelowaniu s-CoMSA

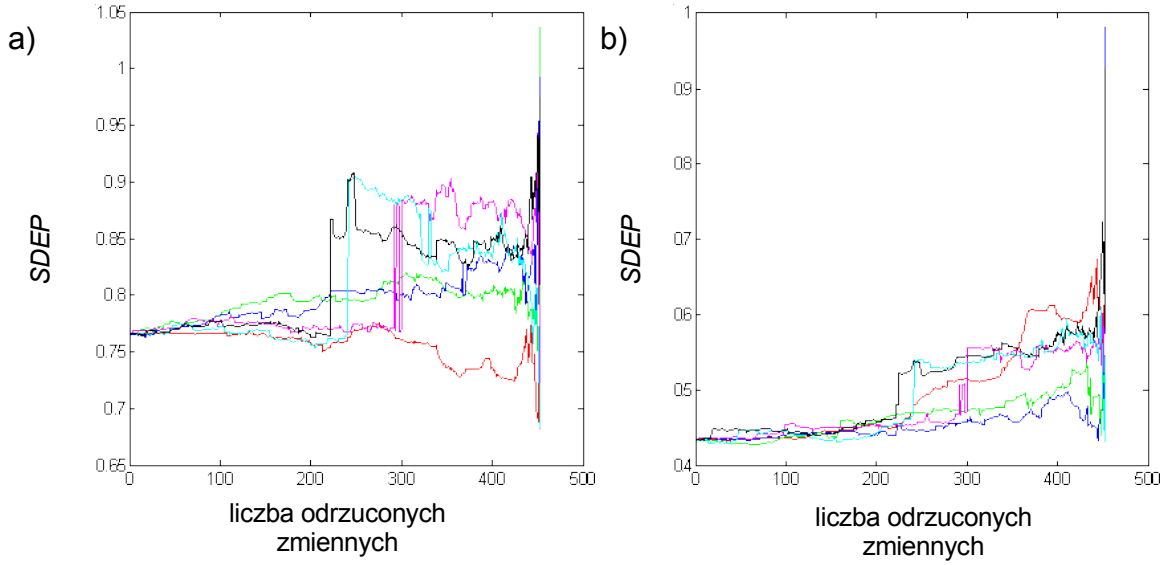
Zastosowanie IVE-PLS w modelowaniu QSAR opisano w publikacji [31]. Jej zaletą jest możliwość swobodnego wyboru liczby wyeliminowanych zmiennych.

Zdolność prognozowania modeli zależy nie tylko od liczby odrzuconych zmiennych ale również od kompleksowości modeli. Rysunek 7.1 przedstawia wykresy zależności parametru q^2_{CV} od liczby wyeliminowanych zmiennych uzyskane dla modelowego zbioru steroidów CBG. Wraz z podwyższaniem założonej *a priori* kompleksowości q^2_{CV} osiąga wyższe wartości. Analogiczny wykres można uzyskać dla parametru *SDEP*. Rysunek 7.2 przedstawia zmianę wartości tego parametru podczas procedury IVE-PLS uzyskaną dla zbioru treningowego steroidów CBG (część a). Widoczna jest na nim zależność parametru *SDEP* od kompleksowości modelu. Większa kompleksowość działa w kierunku wzrostu wartości *SDEP* czyli obniża zdolność prognozowania modelu dla zbioru zewnętrznego. Wykres zależności *SDEP* od rozmiaru sektora, rysunek 5.8b (strona 60), również obrazuje tę tendencję. Obserwowany *SDEP* dla sektora 2 Å wyróżnia się wysoką wartością. Jest to spowodowane wzrostem kompleksowości z 1-2 czynników dla innych rozmiarów sieci do 5 dla sieci 2 Å. Oznacza to, że dobór kompleksowości modelu musi być dokonany bardzo uważnie. W przypadku analizy zbioru modelowego steroidów CBG za optymalny model IVE-PLS został uznany model uzyskany dla założonej maksymalnej kompleksowości 3.

Wyeliminowanie ze zbioru treningowego steroidów CBG związku **s31** powoduje znaczny wzrost zdolności prognozowania modelu dla tego zbioru. Związek **s31** jest postrzegany przez metodę s-CoMSA, podobnie jak przez inne metody 3D-QSAR, jako obiekt odległy [92].



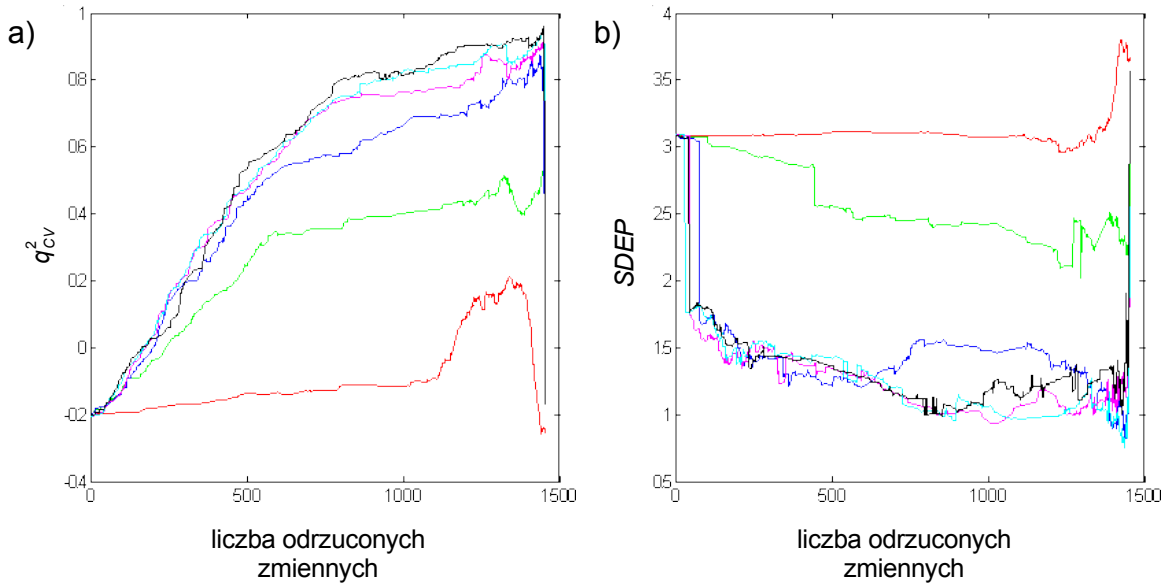
Rysunek 7.1 Wykresy zależności q^2_{cv} od liczby odrzuconych zmiennych podczas procedury IVE-PLS przeprowadzonej na modelowym zbiorze steroidów CBG. Kolorami zaznaczono przebiegi uzyskane dla różnych założonych *a priori* kompleksowości modeli. Kolory w kolejności: czerwony, niebieski, zielony, cyjan, magenta, czarny odpowiadają rozpatrywanej możliwej kompleksowości od 1 do 6.



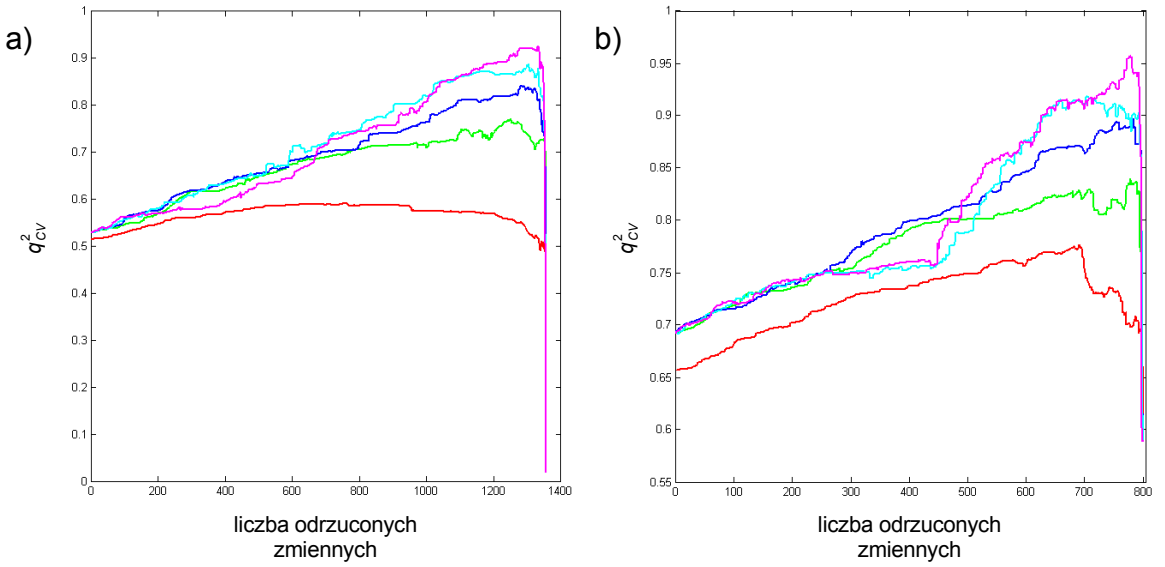
Rysunek 7.2 Wykresy zależności *SDEP* od liczby odrzuconych zmiennych podczas procedury IVE-PLS przeprowadzonej na modelowym zbiorze steroidów CBG. Kolorami zaznaczono przebiegi uzyskane dla różnych założonych kompleksowości modeli. Kolory w kolejności: czerwony, niebieski, zielony, cyjan, magenta, czarny odpowiadają rozpatrywanej możliwej kompleksowości od 1 do 6. Parametr *SDEP* był obliczany dla zbioru testowego (część a) oraz dla zbioru testowego bez związku **s31** (część b).

Odmienne zachowanie zaobserwowano w przypadku procedury IVE-PLS przeprowadzonej na zbiorze modelowym heterocyklicznych barwników azowych nałożonych na siebie w trybie b. Wykresy zależności q^2_{CV} oraz *SDEP* od liczby wyeliminowanych zmiennych są przedstawione na rysunku 7.3. Początkowa wartość q^2_{CV} jest mniejsza od zera. Ujemna wartość parametru q^2_{CV} oznacza, że uzyskany model prognozuje gorzej niż wartość średnia. Początkowa wartość *SDEP* również wskazuje na bardzo słabą zdolność prognozowania. Zastosowanie IVE-PLS znacząco poprawia wartości obu parametrów. Ograniczenie kompleksowości do jednego czynnika powoduje, że procedura IVE-PLS nie poprawia parametrów q^2_{CV} oraz *SDEP*. W pozostałych przypadkach obserwuje się wraz ze wzrostem założonej kompleksowości obniżenie parametru *SDEP*, przy czym pomiędzy kompleksowościami 4, 5 i 6 jego wartość zmienia się tylko nieznacznie.

Podobne efekty zaobserwowano w przypadku szeregu pochodnych α-asaronu oraz inhibitorów NMT – rysunek 7.4.



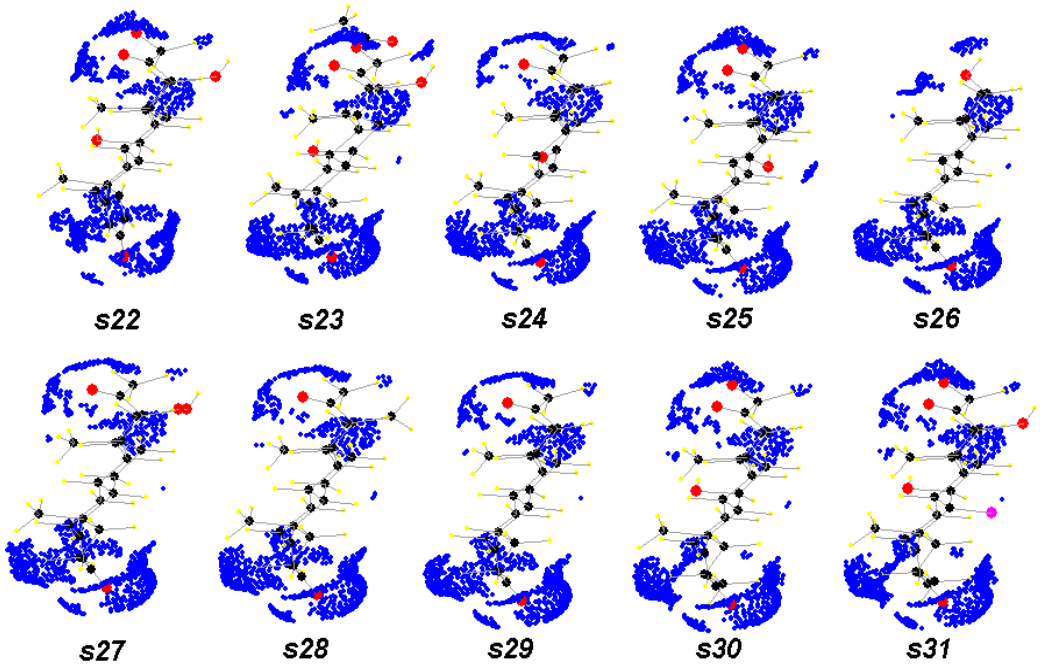
Rysunek 7.3 Wykresy zależności q^2_{CV} (cześć a) oraz *SDEP* (część b) od liczby odrzuconych zmiennych podczas procedury IVE-PLS przeprowadzonej na modelowym zbiorze heterocyklicznych barwników azowych. Kolorami zaznaczono przebiegi uzyskane dla różnych założonych kompleksowości modeli. Kolory w kolejności: czerwony, niebieski, zielony, cyjan, magenta, czarny odpowiadają kompleksowości od 1 do 6. Parametr *SDEP* był obliczany dla zbioru testowego.



Rysunek 7.4 Wykresy zależności q^2_{CV} od liczby odrzuconych zmiennych podczas procedury IVE-PLS przeprowadzonej na zbiorze pochodnych α-asaronu (cześć a) oraz na zbiorze inhibitorów NMT (część b). Kolorami zaznaczono przebiegi uzyskane dla różnych założonych kompleksowości modeli. Kolory w kolejności: czerwony, zielony, niebieski, cyjan, magenta odpowiadają kompleksowościom 2, 4, 6, 8 i 10.

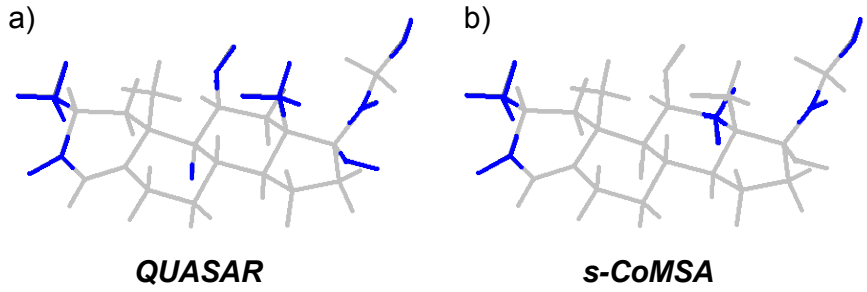
7.1.2 Wizualizacja obszarów oddziaływań specyficznych na podstawie zmiennych typowanych metodą IVE-PLS

Wynikiem działania procedury IVE-PLS jest otrzymanie modelu z mniejszą liczbą zmiennych charakteryzującego się zwiększoną zdolnością prognozowania. Zmienne pozwalające uzyskać lepiej prognozujący model mają spośród całego zbioru zmiennych najistotniejszy wkład w modelowany efekt. W metodzie s-CoMSA każda zmienna odpowiada sektorowi, który z kolei obejmuje ściśle określoną część przestrzeni wokół cząsteczki. Fragmenty powierzchni cząsteczkowych znajdujące się we wskazanych przez IVE-PLS sektorach odpowiadają więc fragmentom mającym najistotniejszy wkład modelowaną aktywność. Przykład tego typu wizualizacji jest przedstawiony na rysunku 7.5. Przedstawione są tam struktury testowego zbioru steroidów CBG. Kolorem niebieskim zaznaczono punkty powierzchni zawarte w sektorach pozostałych po IVE-PLS. Eliminację zmiennych wykonano dla zbioru modelowego, tj. dla związków od **s1** do **s21**, przy maksymalnej możliwej kompleksowości 3. Zmienne pozostałe po eliminacji odpowiadają modelowi, dla którego uzyskano największą wartość q^2_{CV}.



Rysunek 7.5 Najistotniejsze fragmenty oddziaływań zidentyfikowane dla steroidów CBG od **s22** do **s31** na podstawie eliminacji zmiennych IVE-PLS wykonanej dla zbioru modelowego, **s1** do **s21**, przy założonej kompleksowości 3.

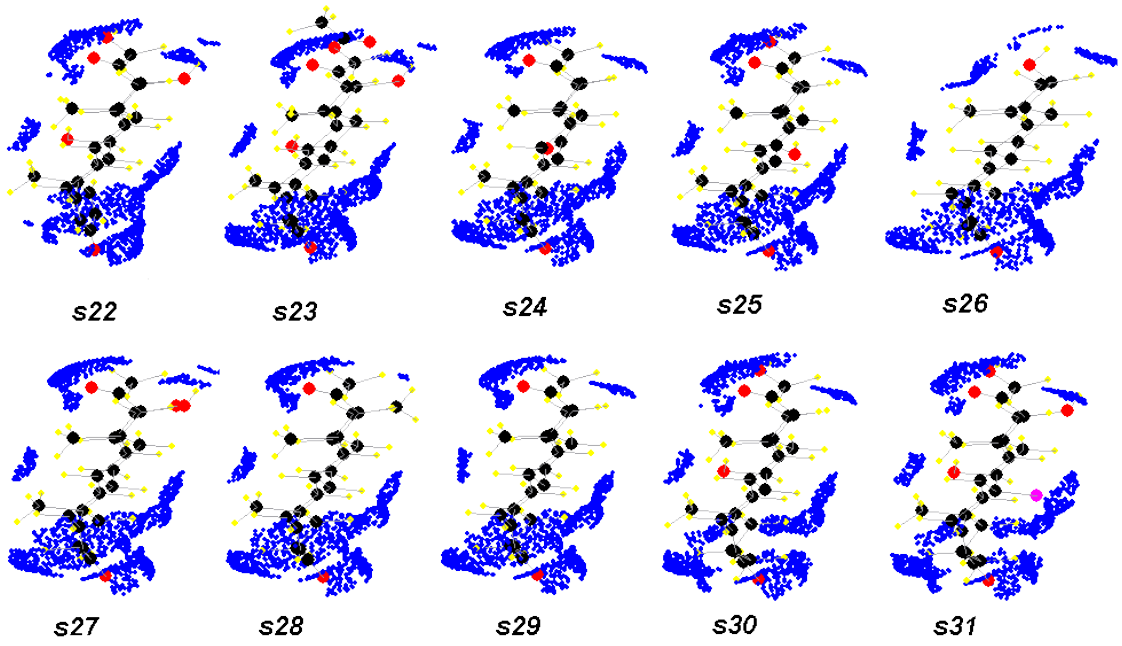
Zidentyfikowane fragmenty powierzchni można również utożsamić z konkretnymi atomami występującymi w analizowanych związkach. Rysunek 7.6a przedstawia fragmenty struktury steroidu CBG **s6** odpowiadające powierzchniom zidentyfikowanym w opisany powyżej sposób. Dla porównania w części b rysunku przedstawiono fragmenty struktury tego samego steroidu zidentyfikowane jako istotne za pomocą oprogramowania Quasar (5D-QSAR) [54]. Można zauważyć, że metoda s-CoMSA wskazuje analogiczne fragmenty.



Rysunek 7.6 Najistotniejsze fragmenty oddziaływań zidentyfikowane dla steroidu CBG **s6** za pomocą metody Quasar (a) [54] oraz na podstawie eliminacji zmiennych IVE-PLS wykonanej dla zbioru modelowego steroidów CBG, **s1** do **s21** (b).

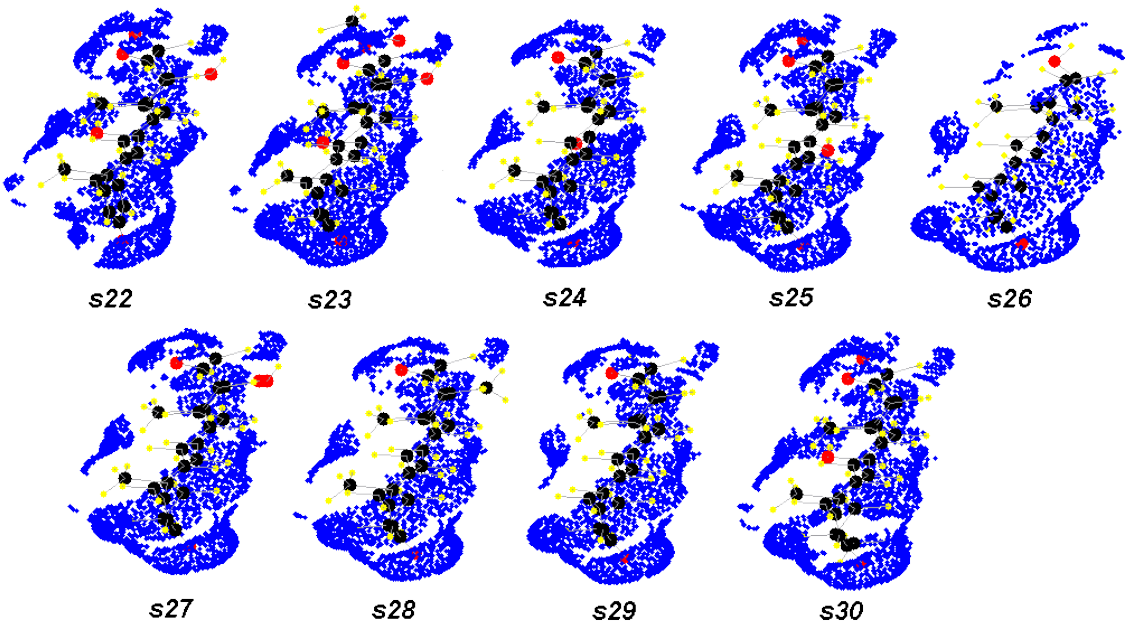
W przypadku steroidów CBG modelowany efekt biologiczny jest powinowactwem cząsteczki do globulin wiążących kortykosteroidy. Zaznaczone na rysunku 7.5 fragmenty można więc uważać za obszary oddziaływań specyficznych między białkiem a cząsteczką steroidów. Znajomość tych obszarów może być pomocna w projektowaniu cząsteczek wykazujących zwiększą aktywność.

Topologia obszarów zidentyfikowanych w opisany powyżej sposób jest uzależniona od przebiegu eliminacji zmiennych. Działanie metody IVE-PLS jest natomiast uzależnione od użytych danych wejściowych. Eliminacja zmiennych przeprowadzona dla tego samego szeregu związków ale dla innego zestawu cząsteczek zmienia przebieg wyboru zmiennych. Rysunek 7.7 przedstawia obszary oddziaływań specyficznych zidentyfikowane za pomocą procedury IVE-PLS wykonanej dla całego zbioru steroidów CBG. Porównanie z rysunkiem 7.5 ujawnia różnice między wskazanymi obszarami. Na rysunku 7.7 wskazany jest mały fragment powierzchni w środkowej części szkieletu steroidowego, którego brak na rysunku 7.5. Występują również różnice we fragmentach w dolnej i w górnej części szkieletu.

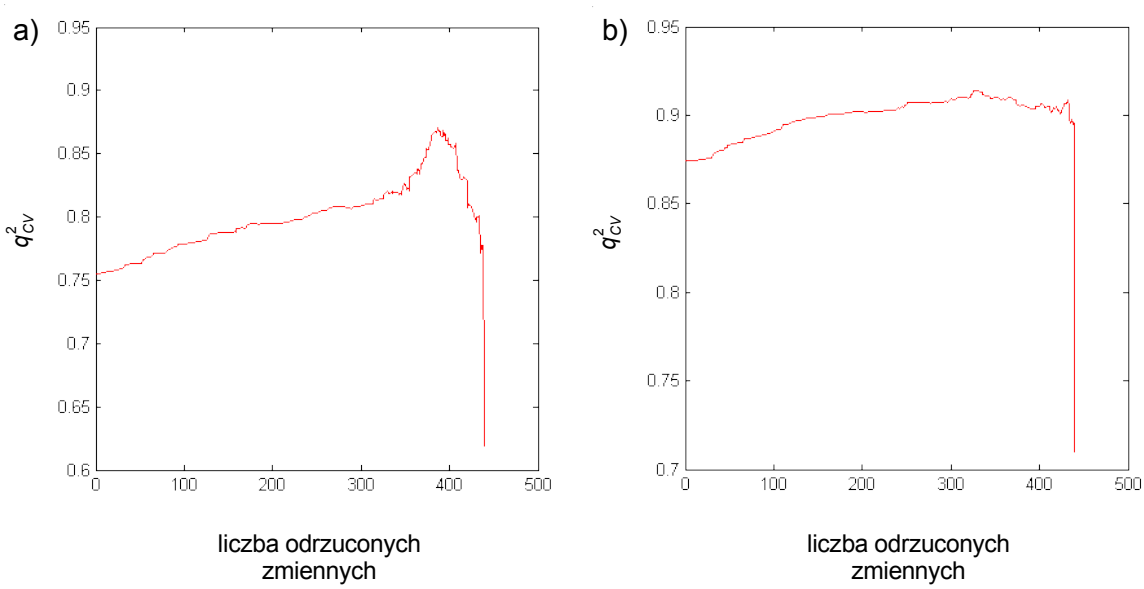


Rysunek 7.7 Najistotniejsze fragmenty oddziaływań zidentyfikowane dla steroidów CBG od **s22** do **s31** na podstawie eliminacji zmiennych IVE-PLS wykonanej dla całego zbioru, **s1** do **s31**.

Rysunek 7.8 przedstawia hipotetyczne obszary oddziaływań receptorowych zidentyfikowane dla zbioru steroidów CBG z pominięciem związku **s31**. Różnice między wskazanymi na nim obszarami a obszarami przedstawionymi na rysunku 7.5 są pod względem jakościowym podobne jak w przypadku rysunku 7.7. Podstawową różnicą między rysunkami 7.7 a 7.8 jest ilość wskazanych obszarów. Związek **s31** nie jest dopasowany do reszty zbioru. Wiele metod 3D-QSAR postrzega go jako obiekt odległy. Uwzględnienie tego związku powoduje wskazanie mniejszej ilości obszarów oddziaływań. Można sądzić, że fragmenty powierzchni wskazane na rysunku 7.7 odpowiadają obszarom oddziaływań wspólnych dla związków **s1** do **s30** oraz dla związku **s31**.



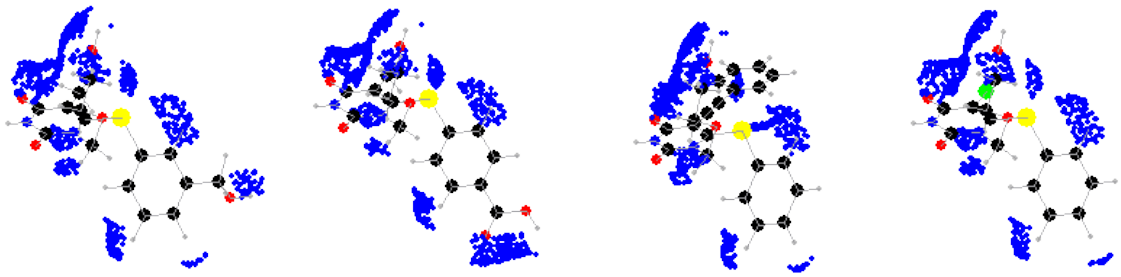
Rysunek 7.8 Najistotniejsze fragmenty oddziaływań zidentyfikowane dla steroidów CBG od **s22** do **s30** na podstawie eliminacji zmiennych IVE-PLS wykonanej dla całego zbioru z pominięciem związku **s31**.



Rysunek 7.9 Wykresy zależności q^2_{cv} od liczby odrzuconych zmiennych podczas procedury IVE-PLS przeprowadzonej na całym zbiorze steroidów CBG (część a) oraz na całym zbiorze z pominięciem związku **s31**.

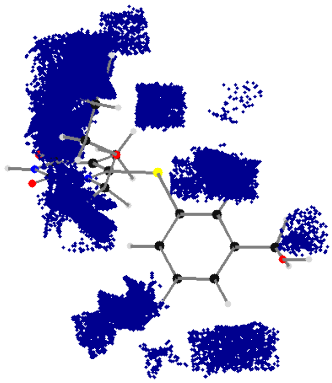
Wykresy zmian wartości q^2_{CV} w trakcie procedury IVE-PLS dla całego zbioru steroidów oraz dla zbioru bez związku **s31** przedstawione na rysunku 7.9 pozwalają na wyjaśnienie różnic na rysunkach 7.7 i 7.8. W przypadku całego zbioru (rysunek 7.7 i 7.9a) początkowa wartość q^2_{CV} jest znacznie niższa a wartość maksymalna jest osiągnięta po wyeliminowaniu niecałych 400 zmiennych. Usuniecie związku **s31** powoduje podniesienie początkowej wartości q^2_{CV}. Wartość maksymalna z kolei jest osiągnięta znacznie wcześniej po wyeliminowaniu około 300 zmiennych (rysunek 7.8 i 7.9b).

Analogiczne obliczenia wykonano dla pochodnych HEPT. Identyfikacja została przeprowadzona dla wszystkich związków szeregu. Znalezione obszary oddziaływań specyficznych są przedstawione na rysunku 7.10 na przykładzie pochodnych **h83**, **h91**, **h97**, **h103**. Większość znalezionych obszarów znajduje się w pobliżu pierścienia heterocyklicznego oraz jego grup bocznych. W pobliżu pierścienia benzenowego również znajdują się wskazane obszary. Pomiędzy obszarami przedstawionymi dla poszczególnych związków występują drobne różnice. W przypadku pochodnej **h91** obszar zlokalizowany w pobliżu pozycji 4 pierścienia benzenowego jest znacznie większy niż w przypadku pozostałych pochodnych. W związku **h83** obszar oddziaływań specyficznych pojawia się pobliżu pozycji 3 pierścienia benzenowego.

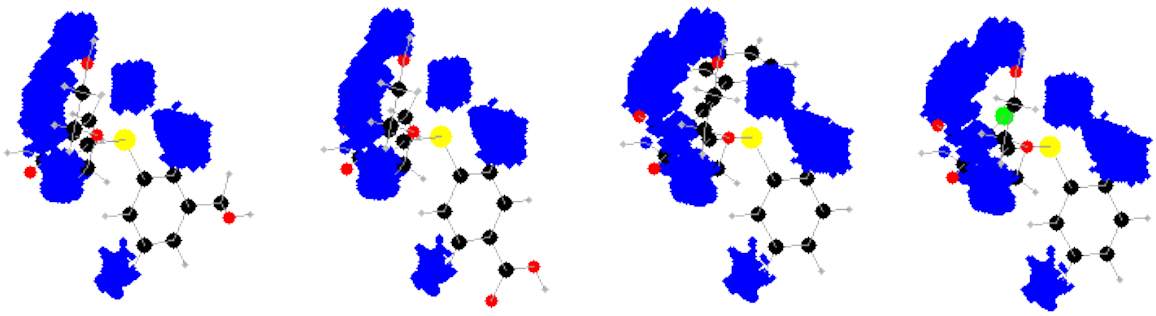


Rysunek 7.10 Najistotniejsze fragmenty oddziaływań zidentyfikowane dla wybranych pochodnych HEPT (związki **h83**, **h91**, **h97**, **h103**) na podstawie eliminacji zmiennych IVE-PLS wykonanej dla całego zbioru

Odmienny obraz oddziaływań specyficznych uzyskuje się przez połączenie fragmentów powierzchni zidentyfikowanych dla wszystkich cząsteczek szeregu. Rysunek 7.11 przedstawia związek **h83** w otoczeniu fragmentów powierzchni będących sumą fragmentów zidentyfikowanych dla wszystkich cząsteczek szeregu. Rysunek 7.12 przedstawia pochodne **h83**, **h91**, **h97**, **h103** w otoczeniu fragmentów uzyskanych przez połączenie obszarów zidentyfikowanych dla wszystkich cząsteczek szeregu z pominięciem 25% najrzadziej występujących obszarów. Uzyskany w ten sposób obrazy są wspólne dla wszystkich cząsteczek analizowanego szeregu [27].



Rysunek 7.11 Najistotniejsze fragmenty oddziaływań zidentyfikowane dla pochodnej **h83** na podstawie eliminacji zmiennych IVE-PLS wykonanej dla całego zbioru. Obraz uzyskany przez połączenie obszarów zidentyfikowanych dla wszystkich cząsteczek szeregu.



Rysunek 7.12 Najistotniejsze fragmenty oddziaływań zidentyfikowane dla wybranych pochodnych HEPT (związki **h83**, **h91**, **h97**, **h103**) na podstawie eliminacji zmiennych IVE-PLS wykonanej dla całego zbioru. Obraz uzyskany przez połączenie obszarów zidentyfikowanych dla wszystkich cząsteczek szeregu z pominięciem 25% najrzadziej występujących obszarów.

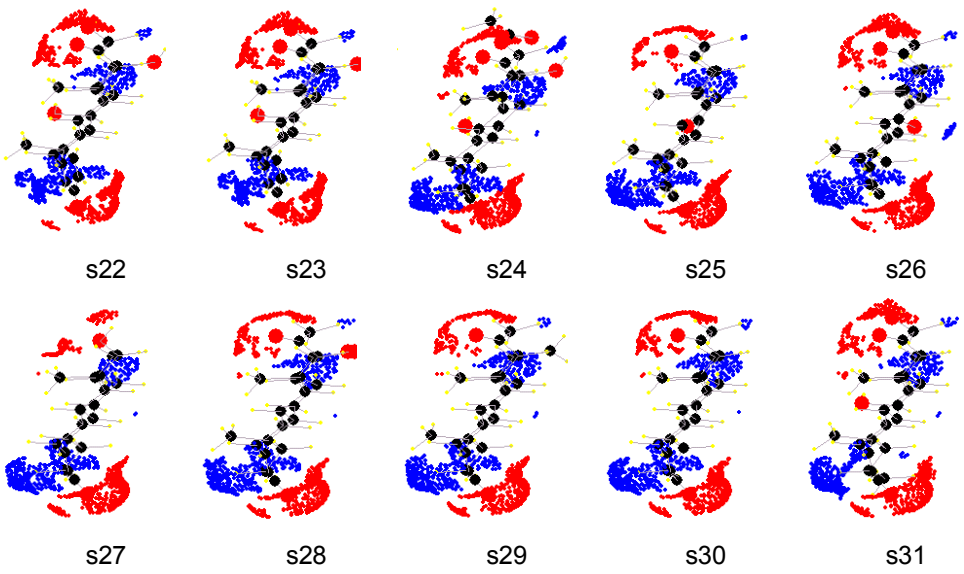
7.2 Ilościowa wizualizacja obszarów oddziaływań specyficznych

Sama identyfikacja obszarów oddziaływań specyficznych bez zaznaczenia ich roli w kształtowaniu badanego efektu utrudnia interpretację modeli. Najprostszym sposobem wizualizacji jest zaznaczenie dodatniego lub ujemnego wkładu zidentyfikowanych fragmentów w modelowaną aktywność. Rysunek 7.13 przedstawia cząsteczki testowego zbioru steroidów CBG w otoczeniu obszarów oddziaływań specyficznych, które dodatkowo oznaczono kolorami czerwonym i niebieskim w zależności od ich dodatniego lub ujemnego wkładu w aktywność. Każdy płat powierzchni odpowiada określonej zmiennej macierzy $\boldsymbol{X}$ zawierającej deskryptor s-CoMSA. Identyfikacja rodzaju wkładu powierzchni jest możliwa po ustaleniu znaku iloczynu wartości zmiennej z odpowiednim współczynnikiem regresji:

$$x_{n\,m} \cdot b_m \tag{7.1}$$

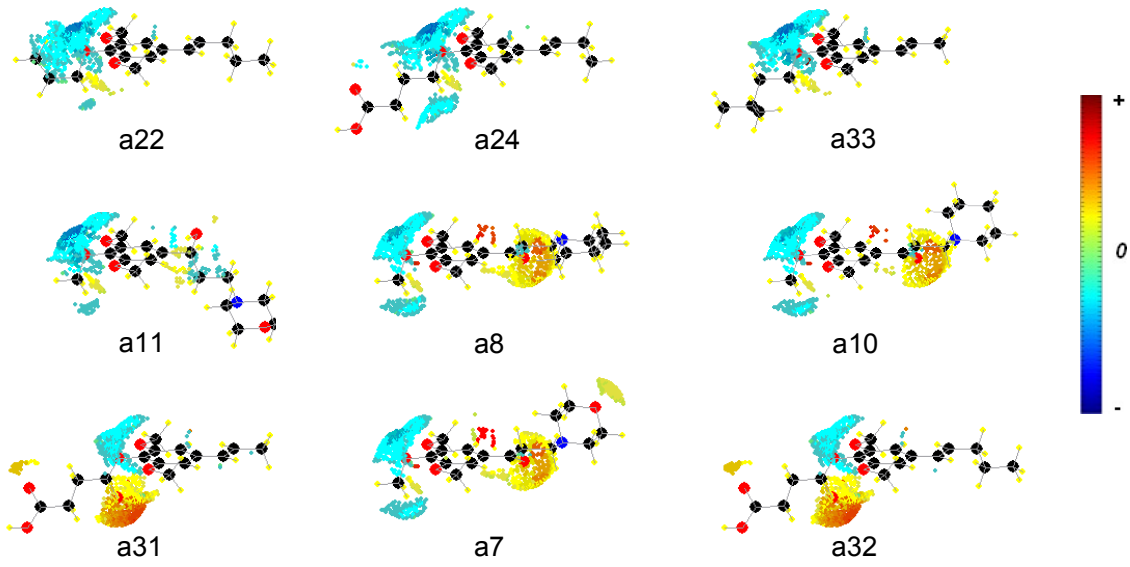
gdzie n indeksuje cząsteczki (obiekty), m indeksuje zmienne, $x_{n\,m}$ oznacza element n m macierzy $\boldsymbol{X}$, b_m jest współczynnikiem regresji zmiennej m. Sens fizyczny wartość zmiennej x jest uzależniony od rodzaju obliczonego deskryptora. W przypadku standardowych analiz s-CoMSA x jest średnią wartością potencjału elektrostatycznego i posiada wymiar tego potencjału. Iloczyn (7.1) jest odpowiednikiem transformacji `Field*Coeff` używanej w wizualizacji modeli CoMFA (patrz również rozdział 4.1, strona 41) [13].

Wartość iloczynu (7.1) może być wykorzystana nie tylko do wizualizacji ale również do detekcji obszarów oddziaływań specyficznych. Rysunek 7.14 przedstawia przedstawia dziewięć wybranych związków z szeregu pochodnych α-asaronu aktywnych w różnym stopniu z zaznaczonymi obszarami oddziaływań specyficznych. Identyfikację tych obszarów wykonano na podstawie modelu PLS uzyskanego dla całego zbioru. Dla każdej zmiennej został obliczony iloczyn (7.1). Następnie odrzucono zmienne, których bezwzględna wartość iloczynu znajdowała się poniżej 90% maksymalnej wartości. W rezultacie otrzymano grupę zmiennych o największym bezwzględnym wkładzie w aktywność.



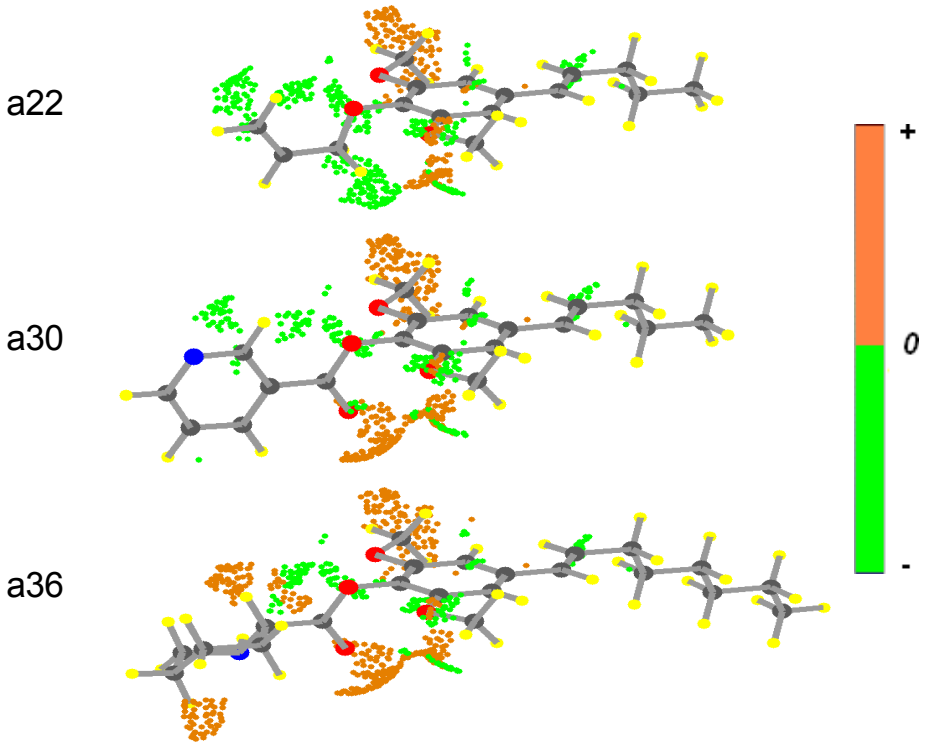
Rysunek 7.13 Najistotniejsze fragmenty oddziaływań zidentyfikowane dla steroidów CBG od **s22** do **s31** na podstawie eliminacji zmiennych IVE-PLS wykonanej dla zbioru modelowego, **s1** do **s21**, przy maksymalnej kompleksowości 3. Pozostałe zmienne odpowiadają modelowi o najwyższym q^2_{CV}. Kolorem czerwonym zaznaczono fragmenty o dodatnim wkładzie w aktywność, niebieskim o wkładzie ujemnym.

Fragmenty powierzchni przedstawione na rysunku 7.14 zostały pokolorowane wartością iloczynu (7.1). Można zauważyć, że obszar znajdujący się na prawo od centralnego pierścienia benzenowego (kolor niebieski) ma ujemny wkład w aktywność co w przypadku szeregu pochodnych α-asaronu jest efektem pożądanym. Obdarzony ładunkiem ujemnym tlen grupy karbonylowej w łańcuchu bocznym, wokół którego znajduje się obszar powierzchni zaznaczony kolorem żółtym i czerwonym powoduje obniżenie aktywności związków – kolory czerwony i żółty oznaczają dodatni wkład w aktywność czyli jej obniżenie. Negatywny wpływ tlenu grupy karbonylowej można zaobserwować w całej grupie związków. Zastąpienie tlenu karbonylowego grupą hydroksylową nie powoduje negatywnego wpływu na aktywność (patrz związki **a10** oraz **a11** na rysunku 7.14). Wpływ tlenu grupy karboksylowej ma kluczowe znaczenie w kształtowaniu aktywności szeregu asaronów. Przykład identyfikacji obszarów oddziaływań specyficznych poprzez progowanie (filtrowanie zmiennych) iloczynu (7.1) pokazuje, że do uzyskania cennych informacji dotyczących wpływu elementów struktury na aktywność związków nie jest konieczna eliminacja zmiennych *explicite*.



Rysunek 7.14 Obszary powierzchni o największym wkładzie na aktywność związków zidentyfikowane na podstawie modelu 2a (tabela 6.6, strona 77). Obszary o bezwzględnym wkładzie w aktywność niższym niż 90% maksymalnej wartości zostały odrzucone. Związki są ułożone w kolejności zmniejszającej się aktywności. Ujemny wkład w aktywność (kolor niebieski) powoduje zwiększenie aktywności. Wkład dodatni (kolor czerwony) powoduje zmniejszenie aktywności.

Identyfikacja obszarów oddziaływań specyficznych asaronów została również wykonana na podstawie zmiennych pozostałych po eliminacji IVE-PLS. Eliminację przeprowadzono dla zbioru z pominięciem związków **a13**, **a29**, **a32** (model 3b tabela 6.6). Rysunek 7.15 przedstawia na przykładzie związków **a22**, **a30**, **a36** zidentyfikowane obszary oddziaływań specyficznych. Kolorem zielonym zaznaczono fragmenty zwiększające aktywność, kolorem pomarańczowym obszary obniżające aktywność. Kluczowe obszary oddziaływań znajdują się w pobliżu centralnego pierścienia aromatycznego. Wpływ łańcucha bocznego na aktywność związków jest wyraźniej zaznaczony. Tlen karboksylowy, podobnie jak w przypadku rysunku 7.14, jest zidentyfikowany jako element obniżający aktywność.



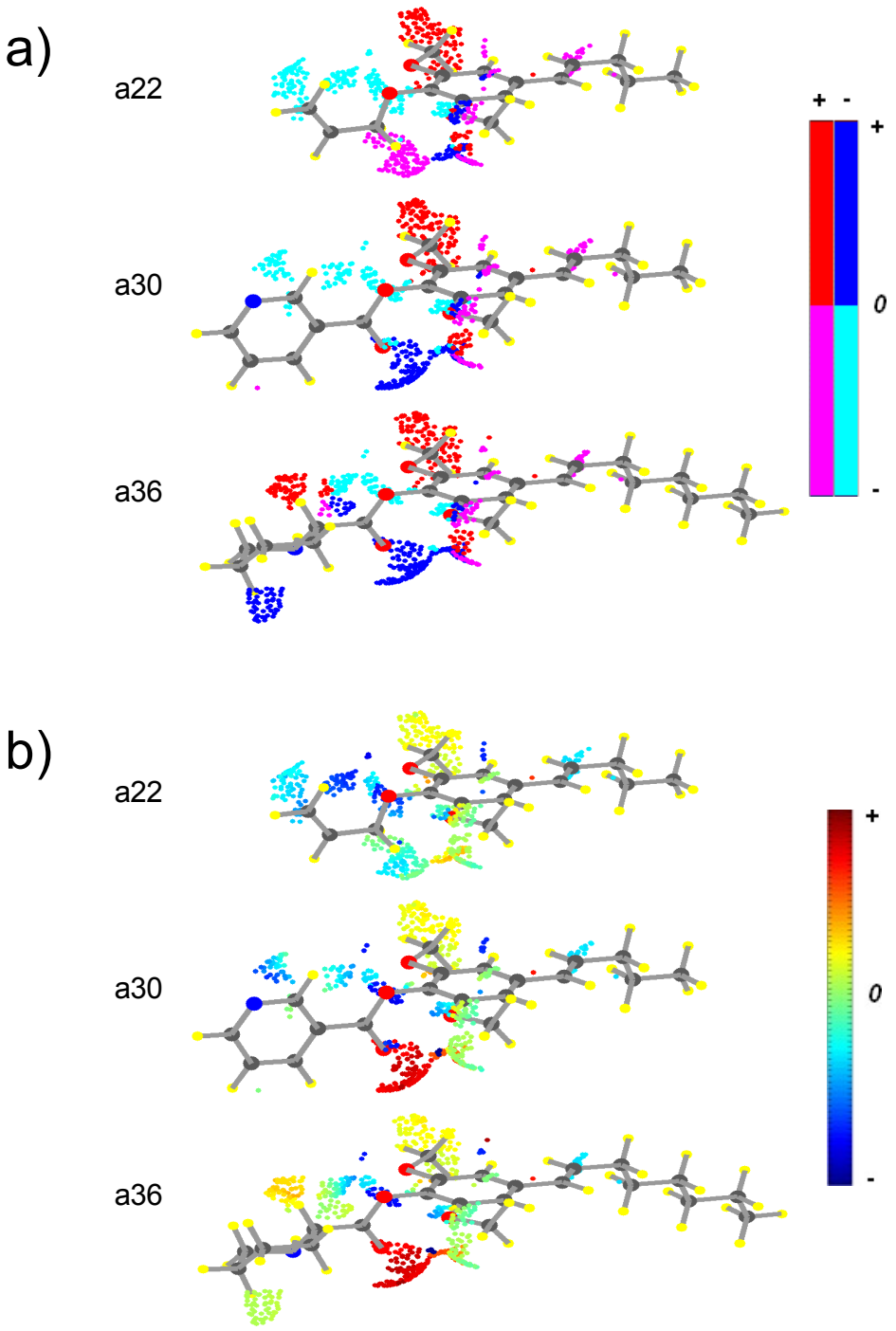
Rysunek 7.15 Obszary oddziaływań specyficznych wybranych pochodnych α-asaronu zidentyfikowane na podstawie modelu IVE-PLS (model 3b – tabela 6.6, strona 77). Ujemny wkład w aktywność (kolor zielony) powoduje zwiększenie aktywność, dodatni wkład w aktywność (kolor pomarańczowy) powoduje zmniejszenie aktywności (patrz również rysunek 7.14).

Te same związki oraz obszary są przedstawione na rysunku 7.16a. W tym przypadku powierzchnie zostały pokolorowane zgodnie ze znakiem czynników iloczynu (7.1). Tabela 7.1 objaśnia znaczenie użytych kolorów:

Tabela 7.1 Znaczenie kolorów używanych w czterokolorowym trybie wizualizacji obszarów oddziaływań specyficznych.

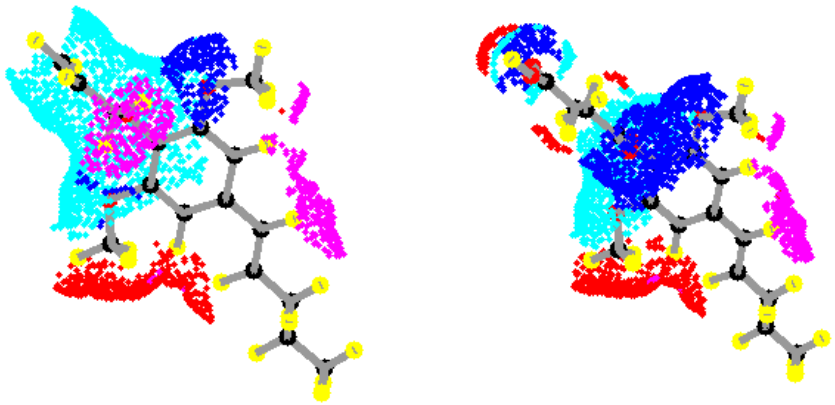
Kolor	Znak czynnika $x_{n\,m}$	Znak czynnika b_m	Wkład w aktywność
czerwony	+	+	dodatni
niebieski	-	-	
magenta	+	-	ujemny
cyjan	-	+	

Na rysunku 7.16b dla porównania obszary oddziaływań specyficznych zostały pokolorowane wartością iloczynu (7.1). Wprowadzenie tlenu karboksylowego do łańcucha bocznego powoduje pojawienie się ujemnego potencjału elektrostatycznego w miejscu, w którym dla związków aktywnych oczekiwany jest potencjał dodatni – kolor niebieski związki **a30**, **a36**. Grupy metoksylowe również są zidentyfikowane jako grupy obniżające aktywność. Rysunek 7.16b pokazuje jednak, że ich wpływ na obniżanie aktywności jest bliski zera (kolor jasno żółty) w przeciwieństwie do wyraźnego wpływu tlenu karboksylowego (kolor ciemno czerwony). Łańcuch boczny połączony z centralnym pierścieniem aromatycznym przez atom tlenu w przypadku związków **a22** i **a30** wprowadza wyraźnie pożądany w tym obszarze ujemny znak potencjału elektrostatycznego.



Rysunek 7.16 Obszary oddziaływań specyficznych dla wybranych pochodnych α-asaronu zidentyfikowane na podstawie modelu IVE-PLS (model 3b – tabela 6.6, strona 77). W części a powierzchnie pokolorowane są zgodnie ze schematem z tabeli 7.1. Kolory niebieski i czerwony oznaczają dodatni udział w aktywności – obniżenie aktywności, kolory cyjan i magenta oznaczają ujemny udział w aktywności – podwyższenie aktywności. W części b te same powierzchnie pokolorowano zgodnie z wartością iloczynu (7.1) – patrz również rysunek 7.14.

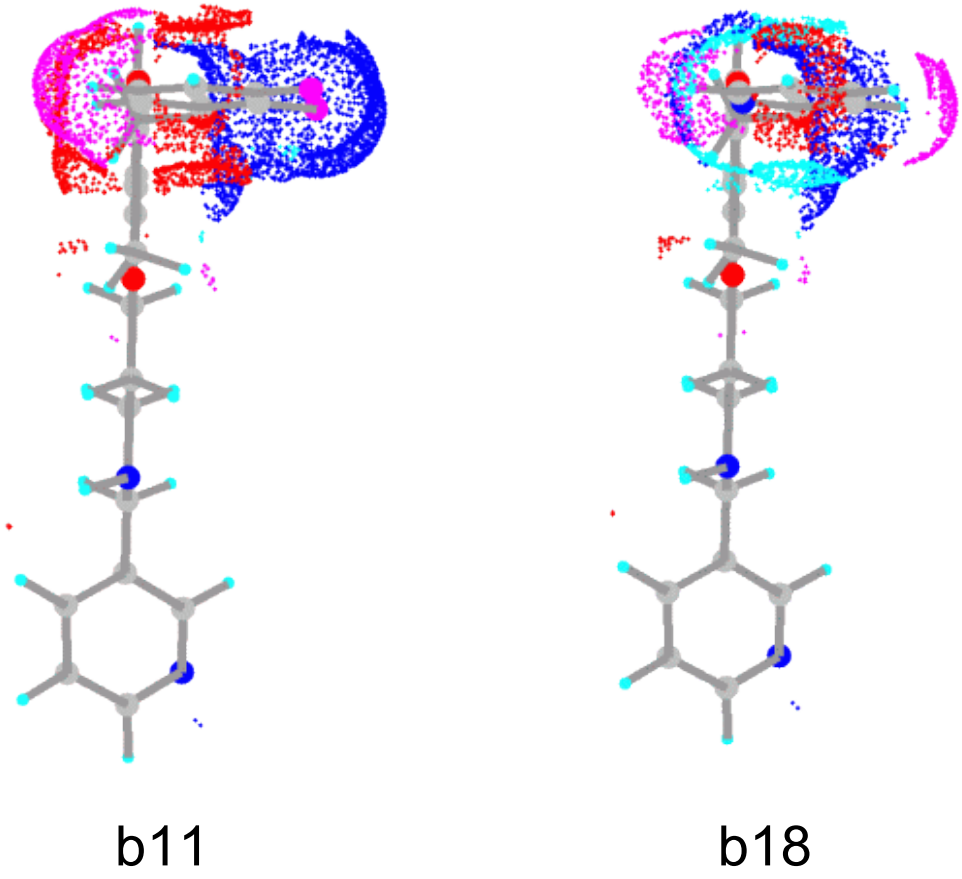
Czterokolorowy schemat wizualizacji został użyty do obszarów zidentyfikowanych przez zastosowanie 90% progu iloczynu (7.1) obliczonego na podstawie modelu PLS uzyskanego dla wszystkich 40 związków szeregu α-asaronów [29]. Rysunek 7.17 przedstawia zidentyfikowane obszary na przykładzie związków **a22** oraz **a32**. Ilość obszarów w porównaniu z rysunkiem 7.14 jest większa. Ponownie potwierdzony jest negatywny wkład w aktywność tlenu karbonylowego w łańcuchu bocznym. Pozytywny wkład powierzchni łańcucha bocznego związku **a22** jest bardziej wyraźny.



Rysunek 7.17 Obszary powierzchni o największym wkładzie w aktywność wybranych pochodnych α-asaronu zidentyfikowane na podstawie modelu 2a (tabela 6.6, strona 77). Obszary o bezwzględnym wkładzie w aktywność niższym niż 90% maksymalnej wartości zostały odrzucone. Powierzchnie zostały pokolorowane zgodnie ze schematem z tabeli 7.1. Kolory niebieski i czerwony oznaczają dodatni wkład w aktywność – obniżenie aktywności, Kolory cyjan i magenta oznaczają ujemny wkład w aktywność – podwyższenie aktywności.

Na rysunku 7.18 przedstawiono obszary oddziaływań specyficznych benzofuranowych inhibitorów NMT. Obszary zostały zidentyfikowane analogicznie jak w przypadku rysunków 7.14 oraz 7.17 na podstawie 90% progu iloczynu (7.1) [29]. Podobnie jak w pozostałych przypadkach kolor niebieski oraz czerwony oznacza dodatni wkład w aktywność – w tym przypadku wiąże się to ze zwiększaniem aktywności. Natomiast obszary w kolorach cyjan i magenta obniżają aktywność. Z opublikowanych badań wynika, że dla aktywności związków istotne są grupy elektronoakceptorowe w pozycjach 2, 3 oraz 5 (patrz rysunek 6.5, strona 79) [124]. Niebieski obszar widoczny dla pochodnej **b11** obrazuje bardzo podobny efekt (rysunek 7.18). Obszar ten oznacza, że dla

aktywnych pochodnych oczekiwana jest w tym miejscu ujemna wartość potencjału. Pochodna **b18** posiada w tym miejscu obszar w kolorze magenta co oznacza występowanie dodatniego potencjału elektrostatycznego.



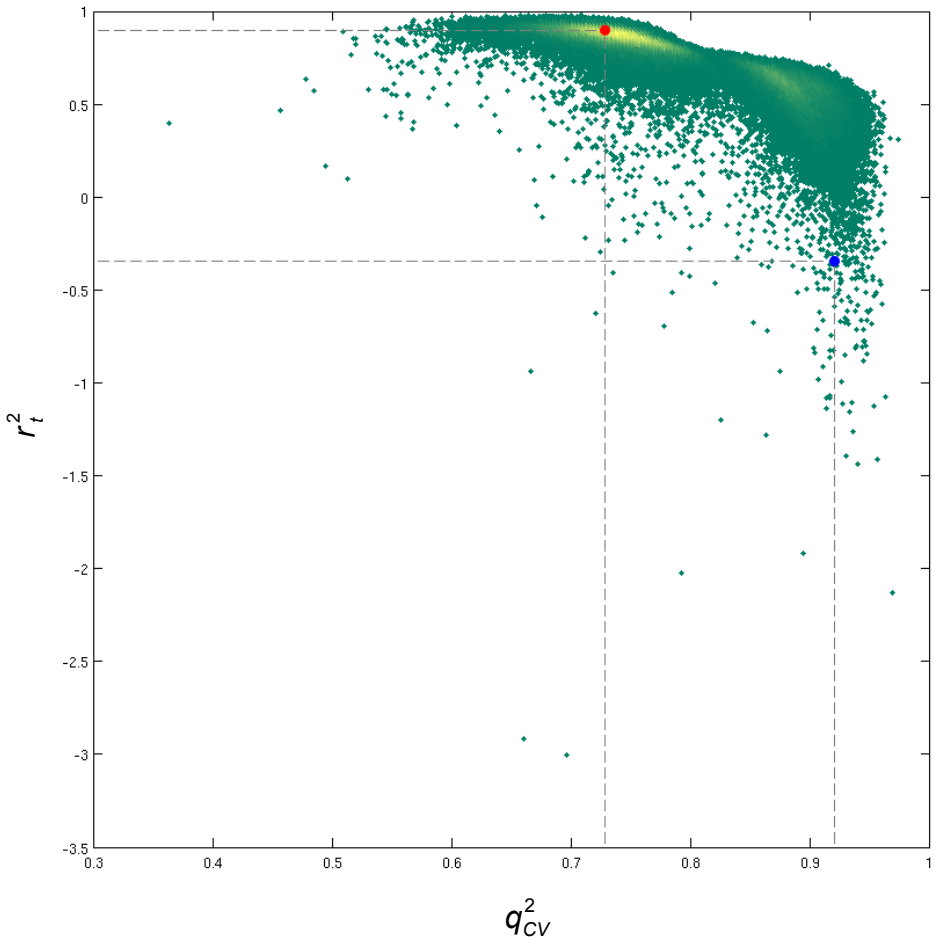
Rysunek 7.18 Obszary powierzchni o największym wkładzie w aktywność wybranych benzofuranowych inhibitorów N-mirystylotransferazy. Obszary o bezwzględnym wkładzie w aktywność niższym niż 90% maksymalnej wartości zostały odrzucone. Powierzchnie zostały pokolorowane zgodnie ze schematem z tabeli 7.1. Kolory niebieski i czerwony oznaczają dodatni udział w aktywności – obniżenie aktywności, kolory cyjan i magenta oznaczają ujemny udział w aktywności – podwyższenie aktywności.

8 Walidacja modeli

W rozdziale 3.4.1 (strona 31) oraz 3.5 (strona 35) omówiono podstawowe metody walidacji modeli stosowane w analizie QSAR. Modele charakteryzujące się dostatecznymi wartościami odpowiednich parametrów są uznawane za wiarygodne. Ocena modelu jedynie na podstawie takich parametrów jak r^2 czy *RMS* jest niewystarczająca, ponieważ do konstrukcji i do testowania modelu stosuje się te same obiekty. Zastosowanie walidacji krzyżowej LSO pozwala na bardziej wiarygodną ocenę modeli. Parametr q^2_{CV} nie może być jednak traktowany jako ostateczna miara jakości modelu. Dopiero użycie całkowicie zewnętrznych zbiorów testowych pozwala na miarodajną walidację modeli [126].

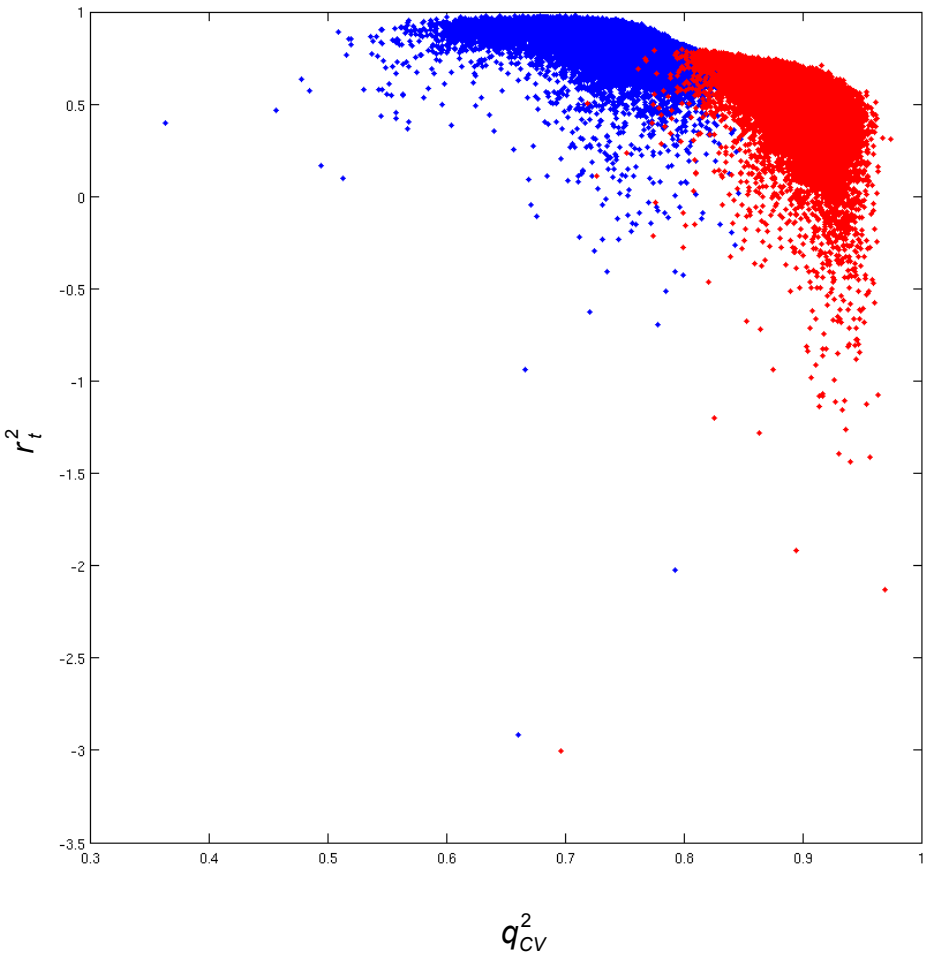
Pomiędzy parametrami wewnętrznej i zewnętrznej walidacji występują pewne korelacje. Idealny przypadek powinien charakteryzować się korelacją dodatnią np. wzrostowi q^2_{CV} powinien towarzyszyć wzrost r^2_t (lub spadek *SDEP*) [127]. Korelacja ujemna oznaczałaby, że wybrana metoda modelowania zawodzi dla danego zbioru danych. W rzeczywistości korelacje występujące między parametrami tego typu są bardzo skomplikowane.

Rysunek 8.1 przedstawia zależność między q^2_{CV} a r^2_t uzyskaną dla steroidów CBG. Każdy punkt wykresu odpowiada podziałowi na zbiór modelowy i testowy. Podział był zawsze wykonywany w proporcjach 21 do 10. Liczba wszystkich możliwych podziałów (44352165) praktycznie uniemożliwia pełną obserwację tego przypadku. Dlatego wybrano co pięćsetny podział – w sumie 88705 podziałów, z których każdy oznaczony jest na rysunku 8.1 (również na rysunku 8.2) jednym punktem. Kolor żółty oznacza wysoką gęstość takich punktów, kolor zielony – odpowiednio niską. Czerwony punkt odpowiada podziałowi uzyskanemu metodą Kennard-Stone [115]. Plasuje się on w obszarze dużej gęstości co oznacza, że podziałów o podobnej charakterystyce jest dużo. W tym zakresie wartość r^2_t jest najwyższa oraz względnie niezależna od q^2_{CV}. Dla wyższych q^2_{CV} widoczna jest tendencja spadkowa r^2_t. Niebieski punkt odpowiada podziałowi literaturowemu. W tym obszarze wartości r^2_t są niższe od 0 co oznacza, że modelowanie całkowicie zawodzi dla zbioru zewnętrznego mimo wysokiej wartości q^2_{CV}.



Rysunek 8.1 Zależność między q_{CV}^2 a r_t^2 uzyskana dla 88705 podziałów zbioru steroidów CBG na zbiór modelowy i testowy w proporcjach 21 do 10. Kolor żółty oznacza rejony o dużej gęstości obserwabli. Kolorem czerwonym zaznaczono punkt odpowiadający podziałowi znalezionemu metodą Kennarda-Stone'a. Kolorem niebieskim zaznaczono punkt opowiadający podziałowi literaturowemu.

Spadek wartości r_t^2 w obszarze wysokich wartości q_{CV}^2 jest spowodowany występowaniem w zbiorze testowym związku **s31**. Związek ten nie jest dopasowany do reszty pochodnych (patrz rozdział 7.1.1, strona 81). Na rysunku 8.2 zaznaczono kolorem czerwonym takie podziały model/test, dla których związek **s31** występuje w zbiorze testowym. Ogromna większość takich przypadków znajduje się w prawej części wykresu w obszarze wysokich wartości q_{CV}^2 i niskich r_t^2. Wyłączenie związku **s31** ze zbioru modelowego pozwala uzyskać wysokie wartości q_{CV}^2 ale jednocześnie jego obecność w zbiorze testowym obniża wartość r_t^2.



Rysunek 8.2 Zależność między q_{CV}^2 a r_t^2 uzyskana dla 88705 podziałów zbioru steroidów CBG na zbiór modelowy i testowy w proporcjach 21 do 10. Kolorem czerwonym zaznaczono występowanie związku **s31** w zbiorze testowym.

8.1 Kryterium Golbraikha – Tropshy

Ciekawą metodę walidacji modeli QSAR zaproponowali Golbraikh i Tropsha [126]. W swojej pracy podkreślają oni, że q^2_{CV} nie może być jedynym kryterium walidacji, oraz że do ostatecznej walidacji należy stosować zewnętrzny zbiór testowy. Wysoka wartość q^2_{CV} (tj. > 0,5) jest kryterium koniecznym, ale niewystarczającym do uzyskania dobrze prognozującego modelu. Zaproponowana przez nich metoda opiera się na założeniu, że wykres zależności między aktywnością zmierzoną a przewidzianą powinien tworzyć linię prostą o korelacji dodatniej. Nachylenie wykresu dla modelu idealnego powinno wynosić 1 a wyraz wolny powinien przyjąć wartość 0. Podobnie, współczynnik korelacji dla takiego modelu powinien być równy 1. Rzeczywisty model o wysokiej wiarygodności powinien być bliski temu obrazowi. Oszacowanie jakości modelu jest dokonywane na podstawie następujących parametrów:

$$k = \frac{\sum_{i=1}^{n} (y_i \; \tilde{y}_i)}{\sum_{i=1}^{n} \tilde{y}_i{}^2} \tag{8.1}$$

$$k' = \frac{\sum_{i=1}^{n} (y_i \; \tilde{y}_i)}{\sum_{i=1}^{n} y_i^2} \tag{8.2}$$

gdzie y_i jest elementem i wektora $\boldsymbol{y}$ zawierającego wartości zmierzonej aktywności, $\tilde{y}_i$ jest elementem i wektora $\tilde{\boldsymbol{y}}$ zawierającego wartości przewidzianej aktywności, k oraz k' oznaczają nachylenie zależności między $\boldsymbol{y}$ a $\tilde{\boldsymbol{y}}$ oraz między $\tilde{\boldsymbol{y}}$ a $\boldsymbol{y}$ liczone bez uwzględnienia wyrazu wolnego, n jest liczbą obiektów zbioru.

Dodatkowo obliczane są współczynniki korelacji:

$$R=\frac{\sum_{i=1}^{n}(y_i-\bar{y}_i)(\tilde{y}_i-\bar{\tilde{y}}_i)}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2\sum_{i=1}^{n}(\tilde{y}_i-\bar{\tilde{y}}_i)^2}} \tag{8.3}$$

$$R_0^2=1-\frac{\sum_{i=1}^{n}(\tilde{y}_i-y_i^{R_0})^2}{\sum_{i=1}^{n}(\tilde{y}_i-\bar{\tilde{y}}_i)^2} \tag{8.4}$$

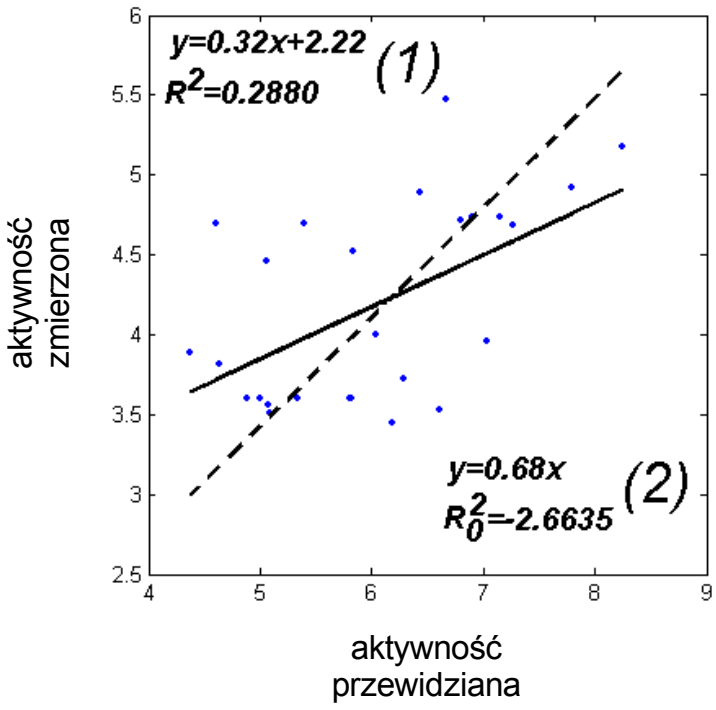
$$R'^2_0=1-\frac{\sum_{i=1}^{n}(y_i-\tilde{y}_i^{R_0})^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2} \tag{8.5}$$

W powyższych wzorach wartości średnie są oznaczane poziomą kreską nad odpowiednim symbolem, y_i jest elementem i wektora $\boldsymbol{y}$ zawierającego wartości zmierzonej aktywności, $\tilde{y}_i$ jest elementem i wektora $\tilde{\boldsymbol{y}}$ zawierającego wartości przewidzianej aktywności, R oznacza współczynnik korelacji, R_0^2 i R'^2_0 są współczynnikami korelacji odpowiadającymi modelom uzyskanym dla k oraz k' gdzie $\boldsymbol{y}^{R_0} = k\,\tilde{\boldsymbol{y}}$ oraz $\tilde{\boldsymbol{y}}^{R_0} = k'\,\boldsymbol{y}$, $y_i^{R_0}$ jest elementem i wektora $\boldsymbol{y}^{R_0}$ a $\tilde{y}_i^{R_0}$ jest elementem i wektora $\tilde{\boldsymbol{y}}^{R_0}$.

Model QSAR jest wiarygodny zgodnie z kryterium Golbraikh i Tropsha (kryterium GT) jeżeli spełnia następujące warunki: $q^2_{CV} > 0{,}5$; $r^2 > 0{,}6$; k lub k' mieszczą się w granicach $< 0{,}85\ 1{,}15 >$ oraz R_0^2 lub R'^2_0 jest bliskie R^2 tj. spełnia warunek [126]:

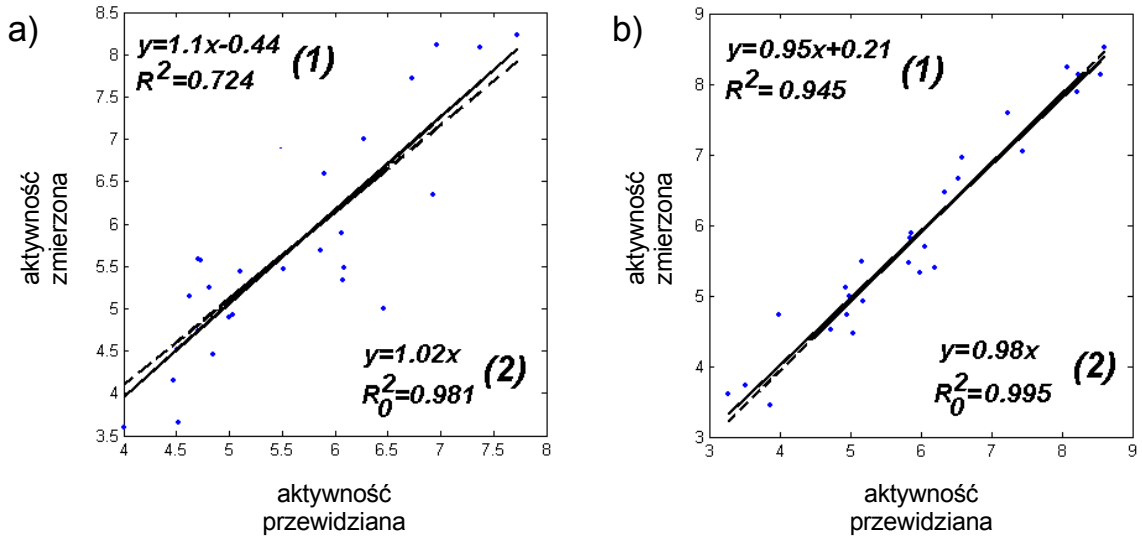
$$\begin{gathered}(R^2-R_0^2)/R^2<0{,}1\\ \vee\\ (R^2-R'^2_0)/R^2<0{,}1\end{gathered} \tag{8.6}$$

Opisane powyżej kryterium zostało zastosowane wobec modelu aktywności pochodnych HEPT. W celu walidacji modelu uzyskanego dla całego szeregu dokonano podziału na zbiory modelowy (związki od **h1** do **h80**) i testowy (związki od **h81** do **h107**). W wyniku takiego podziału uzyskano wysoką wartość *SDEP* 2,01. Dalsza walidacja potwierdziła, że model nie spełnia warunku GT. Warunku tego nie spełnia zresztą także żaden z modeli opublikowanych w literaturze [27, 107, 110]. Rysunek 8.3 przedstawia wykres zależności aktywności zmierzonej od przewidzianej. Linią ciągłą zobrazowano równanie regresji (1), linią przerywaną równanie regresji bez wyrazu wolnego (2). Wartości parametrów R^2, R_0^2 oraz k (0,68) wskazują na bardzo słabą zdolność prognozowania.



Rysunek 8.3 Walidacja modelu uzyskanego dla prostego podziału pochodnych HEPT na zbiór modelowy i testowy za pomocą kryterium Golbraikha-Tropshy. Linia ciągła obrazuje równanie regresji (1), linia przerywana równanie regresji bez wyrazu wolnego (b). Wartości parametrów R^2, R_0^2, k (0,68) nie spełniają ustalonych kryteriów.

Powodem, dla którego modelowanie w zbiorze zewnętrznym zawiodło jest brak reprezentatywności użytych zbiorów. Modelowanie i walidację powtórzono dla zbiorów o tej samej wielkości ale wybranych za pomocą algorytmu Kennard-Stone. Zbiór testowy obejmował związki: **h1**, **h4**, **h5**, **h6**, **h9**, **h10**, **h15**, **h16**, **h23**, **h25**, **h28**, **h29**, **h34**, **h35**, **h37**, **h49**, **h51**, **h55**, **h57**, **h63**, **h64**, **h77**, **h80**, **h81**, **h85**, **h96**, **h103**. Podziałowi temu odpowiada wartość *SDEP* równa 0,69 ($q^2_{CV} = 0{,}72$) [27]. Rysunek 8.4a zawiera wykres zależności aktywności zmierzonej od przewidzianej dla zbioru testowego tego modelu. Podobnie jak poprzednio linią ciągłą zobrazowano równanie regresji (1), linią przerywaną równanie regresji (2). Wartości odpowiednich parametrów spełniają kryterium GT.

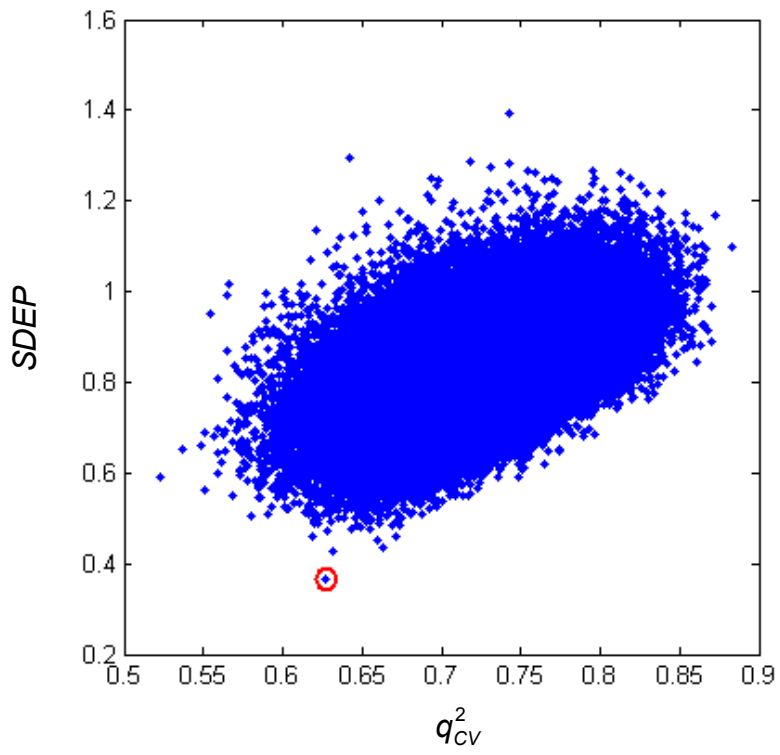


Rysunek 8.4 Walidacja modeli HEPT kryterium Golbraikha-Tropshy. Część a model uzyskany dla podziału metodą Kennarda-Stone'a. Część b model dla najlepszego podziału spośród zbadanych 110000 podziałów (patrz rysunek 8.5). Linie ciągłe obrazują równania regresji (1), linie przerywane równania regresji bez wyrazu wolnego (b). Wartości parametrów R^2, R^2_0, k (część a 1,02; część b 0,98) spełniają ustalone kryteria.

Liczba możliwych do uzyskania podziałów pochodnych HEPT na zbiór modelowy i testowy w proporcji 80 do 27 jest zbyt wielka (około $1{,}5739 \cdot 10^{25}$ podziałów) by móc przeanalizować je systematycznie. Rysunek 8.5 przedstawia wykres zależności między q^2_{CV} a *SDEP* uzyskany dla 110000 losowych podziałów. Podział o najmniejszej wartości *SDEP* (0,37; $q^2_{CV} = 0{,}63$) został przetestowany za pomocą kryterium GT. W zbiorze testowym tego podziału znajdowały się związki: **h3**, **h6**, **h8**, **h11**, **h16**, **h17**, **h22**, **h29**,

h38, **h44**, **h46**, **h52**, **h53**, **h58**, **h60**, **h61**, **h62**, **h64**, **h73**, **h75**, **h76**, **h81**, **h86**, **h87**, **h91**, **h101**, **h103**. Wykres zależności aktywności zmierzonej od przewidzianej dla tego zbioru znajduje się na rysunku 8.4b. Znaleziony w ten sposób podział spełnia kryterium GT [27].

Zależność obserwowana na rysunku 8.5 ma odmienny charakter od przedstawionej na rysunku 8.1 (oraz 8.2) – wykres jest bardziej spójny i jednorodny. Zbiór pochodnych HEPT jest znacznie większy od steroidów CBG, brak też w nim pojedynczych, wyraźnych obiektów odległych. Jednakże podobnie jak w poprzednich przypadkach w zakresie wysokich wartości q^2_{CV} obserwuje się spadek zewnętrznej zdolności prognozowania. W przypadku rysunku 8.5 uwidacznia się to wzrostem wartości *SDEP*. Nadmierny spadek zewnętrznej zdolności prognozowania obserwowany przy wysokich wartościach q^2_{CV} ma związek z przeuczeniem modelu.

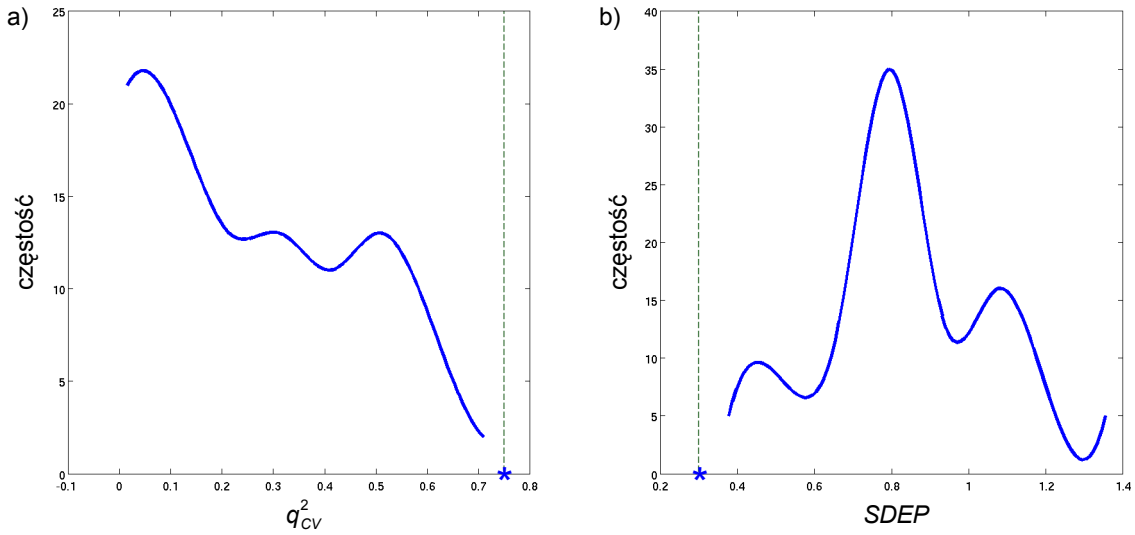


Rysunek 8.5 Zależność między q^2_{CV} a *SDEP* uzyskana dla 110000 losowych podziałów zbioru pochodnych HEPT na zbiór modelowy i testowy w proporcjach 80 do 27. Czerwoną obwódką zaznaczono punkt odpowiadający podziałowi wybranemu do testowania metodą Golbraikha-Tropshy.

8.2 Randomizacja

Przeuczenie modelu jest częstym problemem w analizie QSAR. Model przeuczony charakteryzuje się dobrym przewidywaniem dla zbioru użytego do jego konstrukcji zawodzi natomiast dla zbioru zewnętrznego. Randomizacja jest metodą umożliwiającą wykluczenie przeuczonych modeli. Polega ona na wielokrotnej losowej permutacji wektora zmiennych zależnych. Następnie, dla każdej permutacji tworzony jest odrębny model. W wyniku takiej procedury otrzymuje się histogramy rożnych parametrów oceny zdolności prognozowania [28].

Randomizacja została wykonana w celu przeprowadzenia walidacji modelu aktywności pochodnych α-asaronu. Walidacji został poddany model 6b – patrz tabela 6.6 (strona 77). Permutacja i modelowanie było powtórzone 100 razy. Otrzymane histogramy parametrów q^2_{CV} oraz *SDEP* są przedstawione na rysunku 8.6. Przerywaną linią zaznaczono wartości parametrów odpowiadające oryginalnemu modelowi. Zdecydowana większość modeli posiada niskie wartości q^2_{CV} oraz wysokie wartości *SDEP*. Świadczy to o tym, że model oryginalny jest znacznie bardziej wiarygodny niż modele wygenerowane dla losowych wartości aktywności.



Rysunek 8.6 Histogramy parametrów q^2_{CV} (część a) oraz *SDEP* (część b) uzyskane w wyniku randomizacji modelu aktywności pochodnych α-asaronu. Szczegóły w tekście.

9 Środowisko informatyczne analizy s-CoMSA

Przeprowadzenie obliczeń metodą s-CoMSA wymagało przygotowania odpowiedniego oprogramowania. Metoda została zaprogramowana w środowisku Matlab [128]. Jest to szeroko stosowane środowisko przeznaczone do obliczeń, analizy i wizualizacji danych, ukierunkowane na zastosowania w nauce i przemyśle. Implementacja objęła nie tylko obliczenie samego deskryptora s-CoMSA ale również metody analizy danych QSAR, wizualizację molekuł, obliczenia potencjałów chemicznych, metody identyfikacji obszarów oddziaływań specyficznych, importowanie i eksportowanie danych molekularnych i inne.

9.1 Drug Design Toolbox – DDT

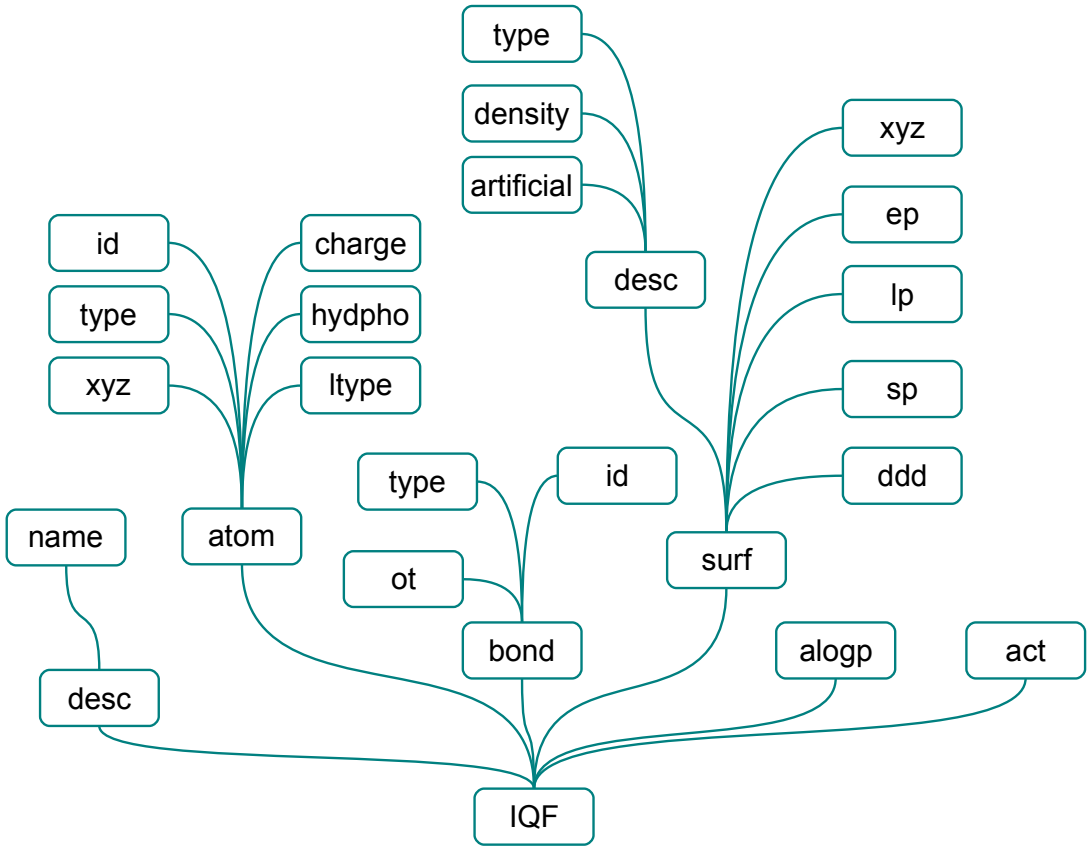
Program został przygotowany w formie pakietu narzędziowego (ang. toolbox) czyli typowej aplikacji środowiska Matlab. Składa się z dwóch warstw. Pierwsza wykonuje wszystkie obliczenia, odpowiada za podstawowe operacje czytania i zapisywania danych. Dostęp do pierwszej warstwy jest możliwy tylko z linii komend okna poleceń programu Matlab. Funkcje stanowiące pierwszą warstwę mogą być z łatwością wywoływane przez skrypty i inne funkcje środowiska Matlab. Drugą warstwę stanowi graficzny interfejs użytkownika. Funkcje drugiej warstwy są odpowiedzialne za wizualną komunikację z użytkownikiem oraz za skomplikowane operacje odczytu i zapisu danych. Wszelkie obliczenia uruchomione w trybie graficznym są wykonywane przez odpowiednie funkcje pierwszej warstwy.

9.2 Formaty danych

Na potrzeby pakietu zostały stworzone dwa formaty danych. Pierwszy format nazwany IQF (ang. internal QSAR format) jest wykorzystywany do zapisu i przechowywania danych molekularnych. Drugi, format UQS (ang. universal QSAR structure), jest przystosowany do przechowywania danych QSAR. Obydwa formaty mogą być zapisywane w postaci binarnych plików `.mat` (standardowe pliki pakietu Matlab) lub w postaci tekstowych plików XML. Dodatkowo pakiet korzysta z własnego formatu prostych baz danych QDB (ang. QSAR data base). Formaty IQF oraz UQS mają strukturą hierarchiczną (strukturę drzewa) i wykorzystują strukturalny typ danych środowiska Matlab.

9.2.1 Internal QSAR format – IQF

Program DDT używa struktur IQF (ang. internal QSAR format) do reprezentacji danych molekularnych. Pojedyncza struktura IQF przechowuje dane o jednej cząsteczce chemicznej. W strukturach IQF zapisywane są miedzy innymi współrzędne i typy atomów, ładunki cząsteczkowe, rodzaje wiązań chemicznych, powierzchnie i potencjały. Rysunek 9.1 przedstawia podstawowy układ danych w strukturze IQF.



Rysunek 9.1 Struktura danych IQF.

Pola **atom** i **bond** zawierają szereg pól potomnych przechowujących dane dotyczące atomów i wiązań. Każdy atom i każde wiązanie posiada własny, unikatowy numer identyfikacyjny zapisany w polach **atom.id** oraz **bond.id**. Pola **atom.type** oraz **bond.type** zawierają odpowiednio typy atomów i wiązań. Pole **atom.xyz** zawiera współrzędne atomów a pole **bond.ot** zawiera identyfikatory atomów tworzących wiązania. Pola **atom.charge**, **atom.ltype** oraz **atom.hydpho** zawierają odpowiednio ładunki cząstkowe, typy atomów zgodne z typami stosowanymi w metodzie ALOGP [129] (metoda kalkulacji Log P patrz rozdział 9.3.2, strona 122) oraz wartości cząstkowej hydrofobowości.

Powierzchnia cząsteczki jest zapisywana w polach gałęzi **surf**. Pole **surf.desc** zawiera podstawowe dane dotyczące powierzchni takie jak typ powierzchni oraz gęstość próbkowania punktów (odpowiednio pola **surf.desc.type** oraz **surf.desc.density**). Znaczenie pól **surf.desc.artificial** oraz **surf.ddd** jest omówione w rozdziale 9.5 (strona 128). Pole **surf.xyz** zawiera współrzędne punktów powierzchni. Potencjał elektrostatyczny, potencjał lipofilowy oraz wartość zawady sterycznej zapisywane są odpowiednio w polach **surf.ep**, **surf.lp** oraz **surf.sp** (patrz rozdział 9.3.1, strona 119).

Pole **alogp** przechowuje wartość Log P obliczonego metodą ALOGP. Pole **act** jest przeznaczone do przechowywania aktywności cząsteczki. Struktura IQF posiada bardzo elastyczną budowę. Większość pól jest opcjonalna. Zależnie od potrzeb mogą być dodawane nowe pola. Pole **act** może być zastąpione innym polem o nazwie bardziej pasującej do przechowywanej wartości, np. **ic50**. Poszczególne gałęzie mogą zawierać również inne pola potomne. Na przykład pole **surf** może zawierać dodatkowe rodzaje potencjałów lub inne wartości charakteryzujące powierzchnię.

9.2.1.1 Typy atomów

W strukturach IQF typ atomu jest zapisywany za pomocą liczby składającej się z 3 pól [`a.hs`]. Pole [`a`] oznacza liczbę atomową pierwiastka, pole [`h`] określa stan hybrydyzacji atomu natomiast pole [`s`] określa dodatkowe właściwości atomu. Prawidłowy typ musi posiadać pole [`a`]. Pozostałe pola są opcjonalne. Tabela 9.1 zawiera opis używanych wartości pola [`h`].

Tabela 9.1 Znaczenie pola [h] w opisie typu atomu stosowanym w strukturach IQF.

Wartość pola [h]	Znaczenie
0	hybrydyzacja atomu jest nieznana
1	hybrydyzacja typu *sp*
2	hybrydyzacja typu sp^2
3	hybrydyzacja typu sp^3
6	atom jest aromatyczny

Pole [s] jest używane tylko do zaznaczenia szczególnych typów atomów. Tabela 9.2 zawiera typy atomów wykorzystujące pole [s].

Tabela 9.2 Typy atomów stosowane w strukturach IQF wykorzystujące pole [s].

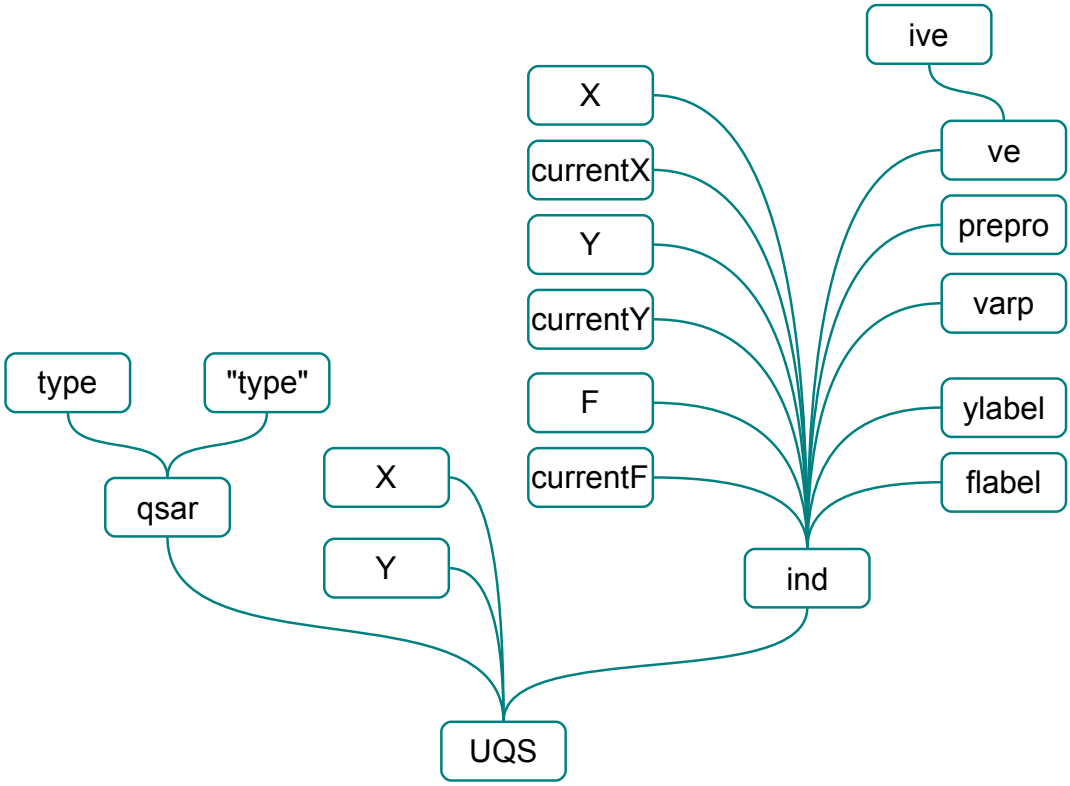
Typ atomu	Opis typu
7.34	atom azotu o hybrydyzacji sp^3 obdarzony ładunkiem dodatnim
7.32	atom azotu o hybrydyzacji sp^3 występujący w wiązaniach amidowych
7.63	aromatyczny atom azotu występujący w pięcioczłonowych pierścieniach heterocyklicznych
7.26	trójwiązalny atom azotu o płaskiej konfiguracji

9.2.1.2 Typy wiązań

Typ wiązania jest w strukturach IQF zapisywany podobnie do typu atomu za pomocą liczby składającej się z 3 pól [b.dt]. Pol [b] oznacza całkowitą krotność wiązania. Pole [d] może przyjmować wartości 0, 5 oraz 6. Cyfra 0 oznacza normalne wiązanie o krotności [b]. Cyfra 5 oznacza wiązanie zdelokalizowane, cyfra 6 wiązanie aromatyczne (w obu przypadkach krotność wiązania wynosi 1). Ostatnie pole [t] jest używane do szczegółowego określenia typu wiązania aromatycznego. W przypadku zwyczajnych wiązań aromatycznych pole [t] ma wartość 5. W przypadku wiązania między hetreroatomem a węglem w pięcioczłonowych heterocyklicznych związkach aromatycznych pole [t] przybiera wartość 4. W ogólności pole [t] jest opcjonalne i znajduje zastosowanie jedynie podczas ustalania cząstkowych hydrofobowości.

9.2.2 Universal QSAR structure – UQS

Struktura UQS (ang. universal QSAR structure) jest używana przez program DDT do zapisu obliczonych deskryptorów QSAR. Pojedyncza struktura UQS przechowuje deskryptory obliczone dla wielu cząsteczek chemicznych przy czym może przechowywać tylko jeden typ deskryptora. Układ danych, podobnie jak w przypadku struktur IQF, ma formę drzewa. Rysunek 9.2 przedstawia podstawowy układ danych w strukturze UQS.



Rysunek 9.2 Struktura danych UQS.

Pole **qsar.type** określa typ użytego deskryptora. W przypadku metody s-CoMSA przyjmuje ono wartość `"scomsa"` a pole **qsar."type"** staje się polem **qsar.scomsa**. Zapisywane są w nim informacje wymagane do obliczenia deskryptora s-CoMSA. Pole **X** przechowuje macierz ***X*** zawierającą wektory deskryptora obliczone dla analizowanych cząsteczek natomiast w polu **Y** znajduje się wektor ***y*** (lub macierz ***Y***) zawierający aktywności. Pole **ind** zawiera szereg pól potomnych używanych do opisu kolumn i wierszy macierzy ***X***, przechowuje ono również wyniki wyboru zmiennych oraz dane używane do identyfikacji oddziaływań specyficznych.

9.2.2.1 Deskryptor s-CoMSA

Struktura UQS jest zaprojektowana do przechowywania deskryptorów dowolnych metod QSAR. Pola **X**, **Y** oraz **ind** są niezależne od metody. Pole **qsar** przechowuje natomiast dane konieczne do wygenerowania deskryptora. W przypadku metody s-CoMSA odpowiednie dane przechowywane są w polu **qsar.scomsa**. Tabela 9.3 zawiera listę pól używanych do generowania deskryptora s-CoMSA oraz ich znaczenie. Wszystkie zawarte w tabeli 9.3 pola są polami potomnymi względem pola **qsar.scomsa**.

Tabela 9.3 Lista pól struktury UQS używanych do generowania deskryptora s-CoMSA.

Pole	Opis
moledge	macierz o wymiarach 2x3, zawiera zakresy wirtualnej siatki na kolejnych osiach współrzędnych
edgeof	wskazuje pole struktur IQF na podstawie, którego obliczany jest deskryptor; zwykle jest to pole **surf.xyz**
molsize	macierz o wymiarach 1x3; zawiera wielkość wirtualnej siatki na kolejnych współrzędnych
celltype	określa typ stosowanego sektora; obecnie jedynym dostępnym typem sektora jest typ cubic oznaczający sektor w kształcie prostopadłościanu
celledge	pole definiuje wielkość sektora; w przypadku sektorów typu `cubic` jest to macierz o wymiarach 1x3 zawierająca wielkości sektora na kolejnych współrzędnych
cellsize	zawiera wartość parametru *cs*, znaczenie tego parametru jest omówione w rozdziale 5.3, strona 52
cs_strict	jeśli równe 1 oznacza, że sektor ma kształt sześcianu foremnego – jest to domyślny kształt sektora
dim	macierz o wymiarach 1x3, określa liczbę sektorów na kolejnych współrzędnych
mesh	macierz o wymiarach Mx3, gdzie M jest liczbą wszystkich sektorów; macierz definiuje położenie poszczególnych sektorów w przestrzeni
mode	wskazuje tryb obliczania wartości sektora, stosowane tryby są omówione w rozdziale 5.2, strona 47
prop	wskazuje pole struktury IQF używane do obliczania wartości sektorów

9.2.2.2 Deskryptor SOM-CoMSA

Obliczanie deskryptora SOM-CoMSA w programie DDT jest możliwe po zainstalowaniu pakietu SOM Toolbox [130]. W przypadku deskryptora SOM-CoMSA pole **qsar.type** przyjmuje wartość "`somcomsa`". Opis deskryptora znajduje się więc w polu **qsar.somcomsa**. Tabela 9.4 zawiera listę pól używanych do wygenerowania deskryptora SOM-CoMSA oraz ich znaczenie. Wszystkie zawarte w tabeli 9.4 pola są polami potomnymi względem pola **qsar.somcomsa**.

Tabela 9.4 Lista pól struktury UQS używanych do generowania deskryptora SOM-CoMSA.

Pole	Opis
sMap	pole przechowuje strukturę sMap generowaną przez SOM Toolbox
md	wartość parametru MD (ang. maximal distance)
iqf	pole przechowuje strukturę IQF związku użytego do treningu sieci neuronowej
mode	wskazuje tryb obliczania wartości sektora, stosowane tryby są identyczne z omówionymi w rozdziale 5.2, strona 47
trainby	wskazuje pole struktur IQF na podstawie, którego trenowana jest sieć neuronowa, zwykle jest to pole **surf.xyz**
prop	wskazuje pole struktury IQF używane do obliczania wartości sektorów

9.2.2.3 Inne deskryptory QSAR

Struktura UQS oraz program DDT są zaprojektowane w sposób umożliwiający łatwe dodanie obsługi innych deskryptorów QSAR.

9.2.2.4 Opis modelu – pole ind

Szczegółowa charakterystyka modelu QSAR jest przechowywana w polu **ind**. Zawiera ono szereg pól potomnych przechowujących informacje o analizowanych cząsteczkach, używanych aktywnościach. Zapisywane są tam wskaźniki używanych zmiennych, aktywności i obiektów a także wyniki wykonanych obliczeń i analiz modelu QSAR.

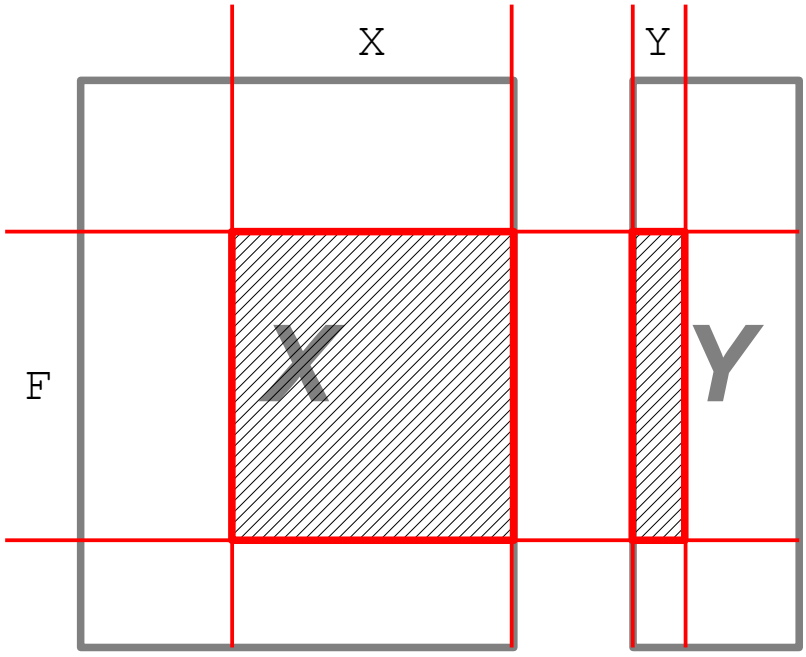
9.2.2.4.1 Opis analizowanych cząsteczek i aktywności

Pole **ind.flabel** przechowuje nazwy wszystkich cząsteczek, których deskryptor jest zapisany w kolejnych wierszach pola **X**. Użyte nazwy powinny być unikatowe.

Podobnie, pole **ind.ylabel**, przechowuje nazwy wszystkich aktywności zapisanych w kolejnych kolumnach pola **Y**.

9.2.2.4.2 Wskaźniki zmiennych, aktywności i obiektów

Struktura UQS może przechowywać dane dla wielu cząsteczek i wielu aktywności. Jednak nie zawsze wszystkie molekuły i nie wszystkie aktywności są wykorzystywane do modelowania. Część molekuł może być pominięta w modelowaniu i używana jako zbiór testowy. Również nie zawsze wszystkie zmienne opisujące cząsteczki są używane do modelowania. Zaznaczenie właściwych zmiennych, aktywności i obiektów jest możliwe odpowiednio za pomocą pól **ind.X**, **ind.Y** oraz **ind.F**. Pola te zawierają bieżące wskaźniki (indeksy) określające aktualnie używane części macierzy $\boldsymbol{X}$ oraz $\boldsymbol{Y}$. Rysunek 9.3 przedstawia schematycznie koncepcję wskaźników w strukturach UQS.



Rysunek 9.3 Wskaźniki zapisane w polu **ind** wskazują aktualnie używane części macierzy ***X*** oraz ***Y***. Indeks `F` (z prawiej strony macierzy ***X***) wskazuje na obiekty obu macierzy a indeksy `X` i `Y` (na górze macierzy ***X*** i ***Y***) na kolumny (zmienne) odpowiednich macierzy. Zakreskowane obszary oznaczają wybrane fragmenty macierzy.

Dodatkowo pole **ind** zawiera trzy pola wskazujące na pochodzenie użytych wskaźników. Są to pola **ind.currentX**, **ind.currentY** oraz **ind.currentF**. Przykładowo, jeżeli w wyniku wstępnego przetwarzania danych część zmiennych zostanie odrzucona to zmienne pozostałe są zapisywane w postaci indeksu `X` w odpowiednim polu, w tym przypadku w polu **ind.prepro.X**. Wówczas, jeżeli w dalszych analizach mają być używane zmienne pozostałe po wstępnym przetwarzaniu indeks `X` z pola **ind.prepro.X** jest kopiowany do pola **ind.X** a pole **ind.currentX** przybiera wartość "`prepro`".

9.2.2.4.3 Zapis wykonanych obliczeń i analiz modelu

Pozostałe pola potomne pola **ind** przechowują wyniki obliczeń i analiz modelu. Każdy rodzaj przeprowadzonej analizy posiada własne pole potomne. Wyniki wstępnego przetwarzania danych są zapisywane w polu **ind.prepro**, wyniki wyboru zmiennych metodą IVE w polu **ind.ve.ive**, wyniki poszukiwania obszarów oddziaływań specyficznych w polu **ind.varp** etc. Jeżeli obliczenia bądź analiza skutkuje uzyskaniem określonego zestawu zmiennych, aktywności lub obiektów odpowiednie wyniki są zapisywane również w postaci wskaźników (patrz rozdział 9.2.2.4.2) w odpowiednich

polach. Dodatkowo wskaźniki obowiązujące w chwili wykonywania obliczeń są zapisywane w polach potomnych względem metody (dla wskaźników `X`, `Y`, `F` odpowiednio w polach **initialX**, **initialY**, **initialF**).

9.2.2.5 Zbiór testowy

Zewnętrzny zbiór testowy może stanowić samodzielna struktura UQS o ile rodzaj stosowanego w niej deskryptora QSAR jest zgodny z deskryptorem zbioru modelowego. Alternatywnie pojedyncza struktura UQS może przechowywać jednocześnie zbiór modelowy oraz zbiór testowy. W tym celu po wygenerowaniu deskryptorów dla cząsteczek obu zbiorów należy w polu **ind.F** wskazać cząsteczki należące do zbioru modelowego. Pozostałe cząsteczki będą należały do zbioru testowego. Moduły programu DDT będą wówczas mogły wykorzystać pole **ind.F** do ustalenia obiektów zbioru modelowego oraz testowego. Odpowiednia opcja, o ile jest dostępna, nosi nazwę `Use reverse F index test data`.

9.2.3 QSAR data base – QDB

Cząsteczki chemiczne są przechowywane pojedynczo w formie struktur IQF. Wykonanie jednej operacji ma wielu cząsteczkach jest możliwe po ich wczytaniu do pamięci lub poprzez bazy danych QDB (ang. QSAR data base). W przypadku dużych zbiorów cząsteczek, rzędu kilku tysięcy struktur, wczytanie całego zbioru do pamięci może niekorzystnie wpłynąć na szybkość obliczeń. Zastosowanie baz danych pozwala wykonywać obliczenia na dużych zbiorach.

Bazy danych QDB są w istocie specjalnie przygotowanym katalogiem zawierającym cząsteczki zapisane w formie struktur IQF lub w innym formacie obsługiwanym przez DDT (Tripos Mol2 [13] oraz CACTVS CTX [131, 132]). Dodatkowo w katalogu będącym bazą QDB musi znajdować się specjalny plik zawierający indeks cząsteczek. Nazwa katalogu, a więc formalnie nazwa bazy QDB, powinna kończyć się znakami „`.qdb`" a plik zawierający indeks cząsteczek musi mieć nazwę „`qdb.mat`" lub „`qdb.qdb`". Program DDT umożliwia tworzenie oraz elementarne zarządzanie bazami QDB zawierającymi cząsteczki w formacie IQF. Tworzenie baz QDB przechowujących pliki Tripos Mol2 lub CACTVS CTX jest możliwe po skopiowaniu odpowiednich plików do pustej bazy i odbudowaniu indeksu cząsteczek (polecenie `File > QDB > Build`).

9.2.4 Importowanie / eksportowanie plików

Pakiet nie jest wyposażony we własny edytor cząsteczek. Nie posiada możliwości optymalizacji geometrii oraz obliczania cząstkowych ładunków atomowych. Optymalizacja geometrii oraz kalkulacja ładunków powinna być wykonana za pomocą zewnętrznych programów. Najszerzej wspieranym zewnętrznym formatem danych molekularnych są pliki w formacie Tripos Mol2 generowane przez program Sybyl [13]. Pakiet DDT potrafi poprawnie odczytywać pliki tego typu przechowujące jedną cząsteczkę. Importuje również poprawnie typy atomów oraz wiązań atomowych co jest konieczne do obliczania Log P metodą ALOGP – patrz rozdział 9.3.2. Eksport do formatu mol2 również przebiega prawidłowo.

Program potrafi dodatkowo czytać pliki w formacie CACTVS CTX [131, 132].

9.3 Właściwości cząsteczkowe

Pierwszym krokiem analizy 3D-QSAR jest odpowiednie przygotowanie zbioru cząsteczek. Programem zalecanym do generowania struktur molekularnych wykorzystywanych przez pakiet DDT jest Tripos Sybyl. Cząsteczki poddawane analizie muszą być prawidłowo narysowane z użyciem prawidłowych typów atomów. Następnie należy wykonać optymalizację geometrii oraz jeżeli w toku dalszej analizy jest planowane obliczanie potencjału elektrostatycznego należy obliczyć cząstkowe ładunki atomowe. Dobrze przygotowany zbiór cząsteczek można łatwo zaimportować do formatu IQF i poddać obliczeniom za pomocą pakietu DDT.

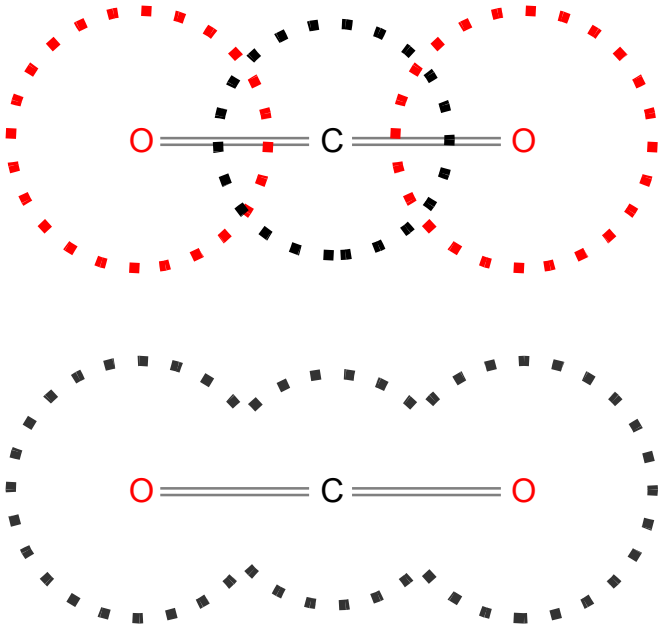
9.3.1 Powierzchnie

Program DDT posiada możliwość generowania powierzchni cząsteczkowych w postaci zbioru punktów z powierzchni. Algorytm generowania powierzchni składa się z dwóch etapów. W pierwszym etapie wokół wszystkich atomów generowane są punkty leżące na sferach o promieniach van der Waalsa odpowiednich atomów. Gęstość wygenerowanych punktów może być dowolna, domyślna gęstość wynosi 25 punktów na $Å^2$. W drugim etapie usuwane są punkty znajdujące się wewnątrz sąsiednich sfer. W rezultacie otrzymywana jest prosta powierzchnia van der Waalsa. Rysunek 9.4 przedstawia schematycznie opisany sposób generowania powierzchni.

Po wygenerowaniu powierzchni w opisany powyżej sposób w każdym wygenerowanym punkcie możliwe jest obliczenie kilku potencjałów chemicznych. Jeżeli cząsteczki zostały zaimportowane wraz z ładunkami cząstkowymi możliwe jest obliczenie potencjału elektrostatycznego. Program umożliwia również kalkulację potencjału lipofilowego (patrz rozdział 9.3.2). Dodatkowo możliwe jest obliczenie wartości tzw. zawady sterycznej. Jest to własność obrazująca dostępność powierzchni w danym punkcie. Jest obliczana w podobny sposób jak pole oddziaływania sterycznego w metodzie CoMSIA [17]:

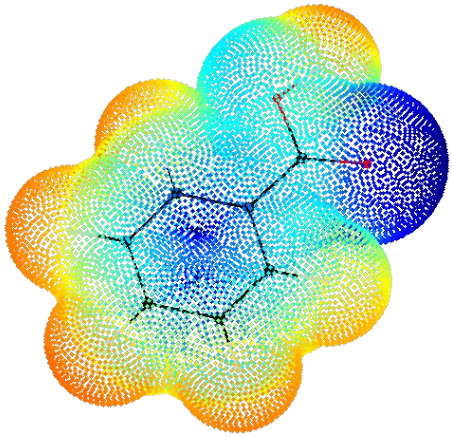
$$SP_k = \sum_i \frac{v_i}{p_{ik}} \tag{9.1}$$

gdzie SP_k jest zawadą steryczną w punkcie k, v_i jest promieniem van der Waalsa atomu i, p_{ik} jest miarą odległości obliczoną za pomocą wzoru (9.3) – strona 122.

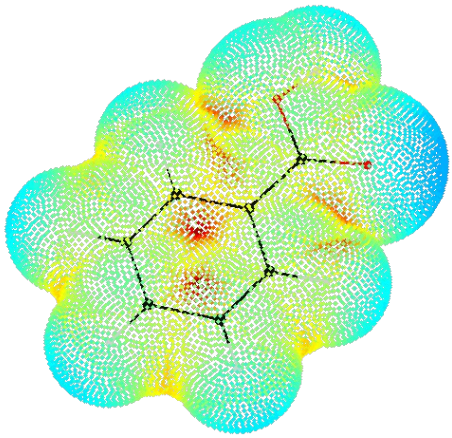


Rysunek 9.4 Schemat generowania powierzchni van der Waalsa.

Rysunki 9.5 i 9.6 przedstawiają powierzchnię kwasu karboksylowego wygenerowaną w programie DDT pokolorowaną potencjałem elektrostatycznym i wartością zawady sterycznej. Wodory w pozycji orto w przypadku rysunku 9.6 nie różnią się znacznie od pozostałych wodorów. Właściwość powierzchni obliczana za pomocą wzoru (9.1) ma charakter ściśle lokalny. Wpływ atomów znajdujących się w odległości przekraczającej 1-2 Å nie jest uwzględniany.



Rysunek 9.5 Powierzchnia kwasu benzoesowego pokolorowana wartością potencjału elektrostatycznego.



Rysunek 9.6 Powierzchnia kwasu benzoesowego pokolorowana wartością zawady sterycznej obliczoną za pomocą wzoru (9.1).

9.3.2 Log P, potencjał lipofilowy

Jedną z wielu metod szacowania Log P jest opracowana przez Ghose i Crippena metoda ALOGP [129]. Jest to stosunkowo prosta metoda umożliwiająca dobre oszacowanie Log P. Działanie metody ALOGP polega na przypisaniu każdemu atomowi zależnie od jego typu oraz od topologii połączeń z innymi atomami jeden ze 120 stabelaryzowanych typów atomowych. Każdemu nowemu typowi atomowemu odpowiada określona tzw. cząstkową hydrofobowość. Tabela cząstkowych hydrofobowości została ustalona eksperymentalnie na drodze pomiaru Log P przeprowadzonego dla około 9000 związków chemicznych. Wartość Log P w metodzie ALOGP jest obliczana przez zsumowanie cząstkowych hydrofobowości wszystkich atomów.

Log P jest makroskopową własnością cząsteczek chemicznych. Potencjał lipofilowy, jest natomiast własnością mikroskopową, którą należy uważać za lokalne powinowactwo do cząsteczek wody. Można go obliczyć na podstawie cząstkowych hydrofobowości [133]. Wartość Log P nie posiada wymiaru, cząstkowe hydrofobowości otrzymane metodą ALOGP są również bezwymiarowe. Przez to również obliczony potencjał lipofilowy nie posiada wymiaru. Program DDT do kalkulacji potencjału lipofilowego wykorzystuje wzór:

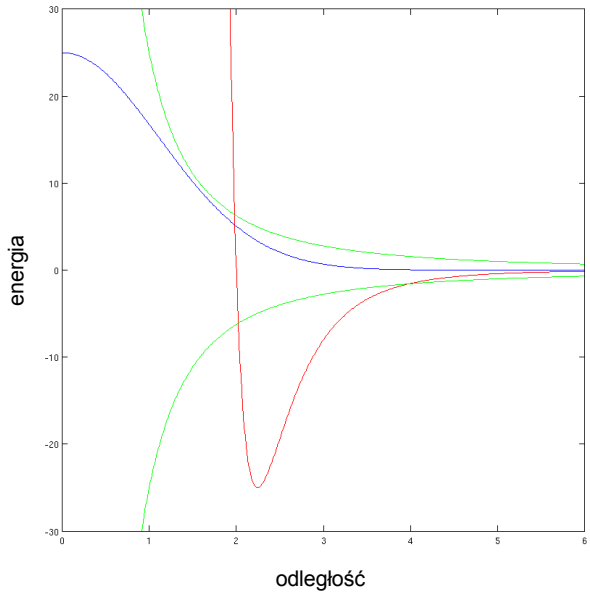
$$LP_k = \sum_i \frac{f_i}{(1 + p_{ik}/10)} \tag{9.2}$$

Jest to zmodyfikowany wzór z publikacji [133]. W powyższym wzorze LP_k oznacza wartość potencjału lipofilowego w punkcie k, i indeksuje atomy, f_i jest cząstkową hydrofobowością atomu i, p_{ik} jest odpowiednikiem odległości między atomem i a punktem k obliczonym za pomocą wzoru (9.3):

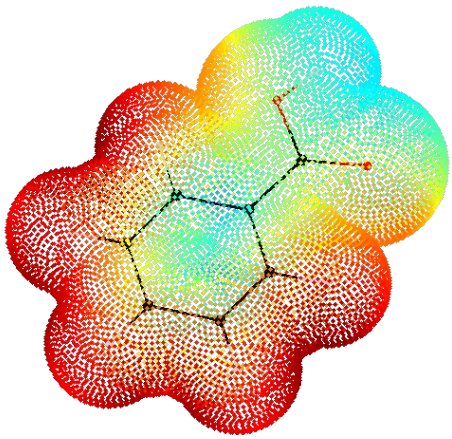
$$p_{ik} = e^{\alpha \cdot r^2} \tag{9.3}$$

gdzie r jest odległością kartezjańską między punktami i oraz k, α jest stałym czynnikiem wyciszającym o wartości 0,4. Wzór (9.3) wyraża w istocie potencjał oddziaływania stosowany w metodzie CoMSIA. Rysunek 9.7 przedstawia wykresy wybranych potencjałów chemicznych. Potencjał stosowany w metodzie CoMSIA szybko tłumi dalsze oddziaływania a dla oddziaływań bliskich nie zmierza do nieskończoności lecz przybiera ustaloną wartość. Czynnik α przyjmuje zwykle wartość z zakresu od 0,2 do 0,4. Większe wartości powodują zwiększenie nachylenia potencjału oraz powodują mocniejsze

tłumienie dalszych oddziaływań. Na potrzeby kalkulacji potencjału lipofilowego przyjęto wartość α 0,4 by podkreślić lokalny charakter obliczanej właściwości. Przykładowa powierzchnia pokolorowana potencjałem lipofilowym jest przedstawiona na rysunku 9.8.



Rysunek 9.7 Potencjały stosowane w chemii. Potencjał Lennarda Jonesa – kolor czerwony, potencjał Coulomba – kolor zielony, potencjał funkcji Gaussa stosowany w metodzie CoMSIA – kolor niebieski.

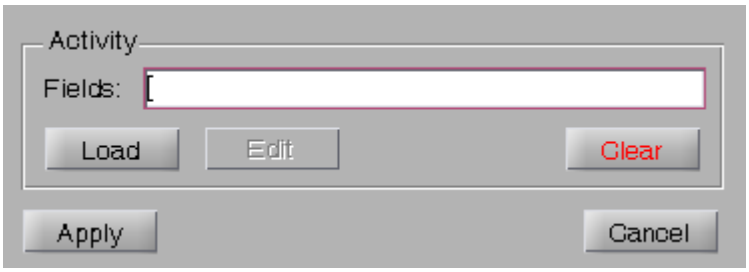


Rysunek 9.8 Powierzchnia kwasu benzoesowego pokolorowana wartością potencjału lipofilowego obliczonego za pomocą wzoru (9.2).

9.3.3 Aktywność

W rozdziale 9.2.1 opisano pola struktur IQF przeznaczone do przechowywania aktywności cząsteczek. W celu zapisania aktywności w strukturach IQF należy najpierw w środowisku Matlab wprowadzić je do macierzy a macierz zapisać w pliku `.mat`. Następnie w oknie `IQF activity fields` (`Molecules > Manage, Activity`) w polu `Fields` należy wprowadzić oddzielone przecinkiem nazwy pól, w których mają być zapisane aktywności. Przycisk `Load` umożliwia wczytanie zapisanych wcześniej aktywności.

Aktywności zapisane w odpowiednich polach struktur IQF są wykorzystywane w dalszych obliczeniach przeprowadzanych w programie DDT.



Rysunek 9.9 Okno wprowadzania aktywności.

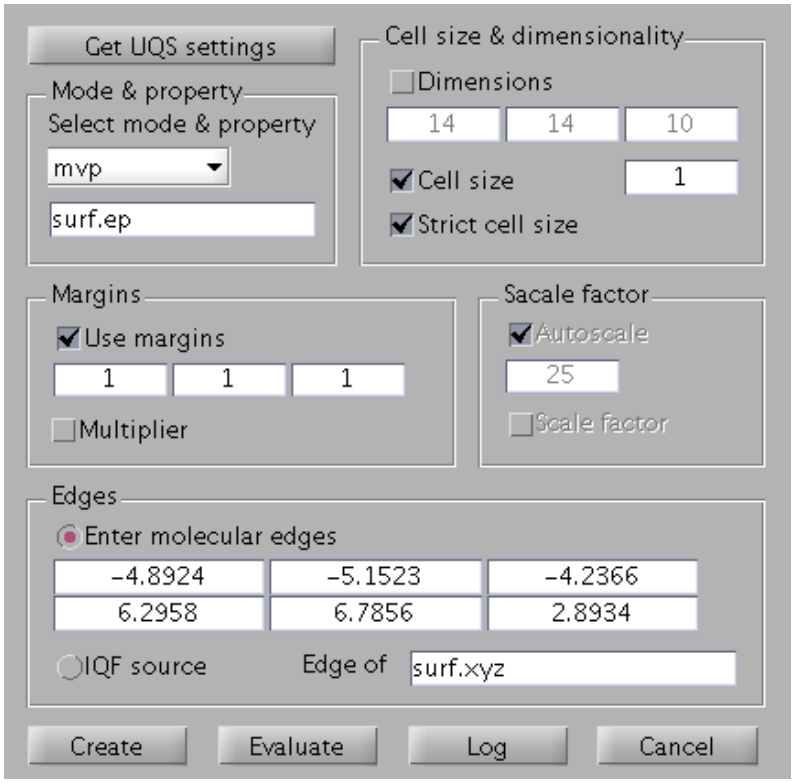
9.4 Generowanie deskryptorów QSAR

Obliczanie deskryptorów QSAR w programie DDT jest podzielone na dwa etapy. W pierwszym etapie ustawiane są wszystkie konieczne opcje i tworzony jest początkowy plik UQS. Drugi etap polega na generowaniu deskryptorów dla wskazanych cząsteczek (zgodnie z ustawieniami z pierwszego etapu) i zapisywaniu ich we skazanym pliku UQS.

9.4.1 Generowanie deskryptora s-CoMSA

Opis deskryptora s-CoMSA od strony formalnej znajduje się w rozdziale 5.1. Posługiwanie się takim formalizmem jest niewygodne z informatycznego punktu widzenia. Program DDT posiada moduł pozwalający ustawić wszelkie opcje konieczne do obliczania deskryptora. W rozdziale 9.2.2.1 omówiono sposób zapisu ustawień s-CoMSA w strukturach UQS. Za pomocą modułu deskryptora s-CoMSA wprowadzane są wartości pól wymienionych w tabeli 9.3 lub dane służące do ich obliczenia.

Rysunek 9.10 przedstawia okno ustawień deskryptora s-CoMSA. W ramce `Mode & property` można wybrać tryb obliczania deskryptora: `mvp`, `nps`, `mvp/nps` (patrz wzory 5.3, 5.4 oraz 5.5 – strona 49). Jeżeli wybrano tryb `nps` lub `mvp/nps` istnieje możliwość podania czynnika skalującego obliczane wartości (opcja `Scale factor`). Program może również obliczyć czynnik skalujący automatycznie na podstawie gęstości stosowanej powierzchni (opcja `Autoscale`). Ramka `Mode & property` zawiera również okienko, w którym należy podać własność, w formie pola struktury IQF, używaną do obliczania wartości sektora (wymagane w przypadku trybów `mvp` oraz `mvp/nps`). W przypadku potencjału elektrostatycznego należy wprowadzić ciąg znaków: `surf.ep`. Wybrana własność musi być zgodna z polem `Edge of` z ramki `Edges`, które standardowo wskazuje pole przechowujące powierzchnię. W tej samej ramce znajduje się sześć okienek z zakresami wirtualnej siatki na trzech kolejnych osiach układu. Wartości te mogą być wpisane ręcznie lub obliczone automatycznie na podstawie plików IQF – opcja



Rysunek 9.10 Okno ustawień deskryptora s-CoMSA programu DDT.

`IQF source`.

Ramka `Margins` umożliwia podanie dodatkowych marginesów siatki. Marginesy mogą być podane jako liczby wyrażona w Å dodawane do rozmiarów siatki lub jako mnożniki. Ramka `Cell size & dimensionality` umożliwia ustawienie rozmiaru komórki siatki. Rozmiar może być ustalony przez wprowadzenie liczby komórek mieszczących się na kolejnych osiach układu (opcja `Dimensions`) lub przez podanie rozmiaru komórki (opcja `Cell size`). W pierwszym przypadku rozmiar komórki będzie zależny od wymiarów całej siatki oraz od wprowadzonej liczby komórek mieszczących się na kolejnych osiach. W drugim przypadku od podanego rozmiaru komórki będzie zależała liczba komórek na kolejnych osiach. Opcja `Stric cell size` powoduje, że wszystkie wymiary komórki są sobie równe tj. sektory są sześcianami foremnymi. Powoduje to konieczność korekty ustalonych rozmiarów siatki (korekta odbywa się automatycznie, bez udziału użytkownika).

9.4.2 Generowanie deskryptora SOM-CoMSA

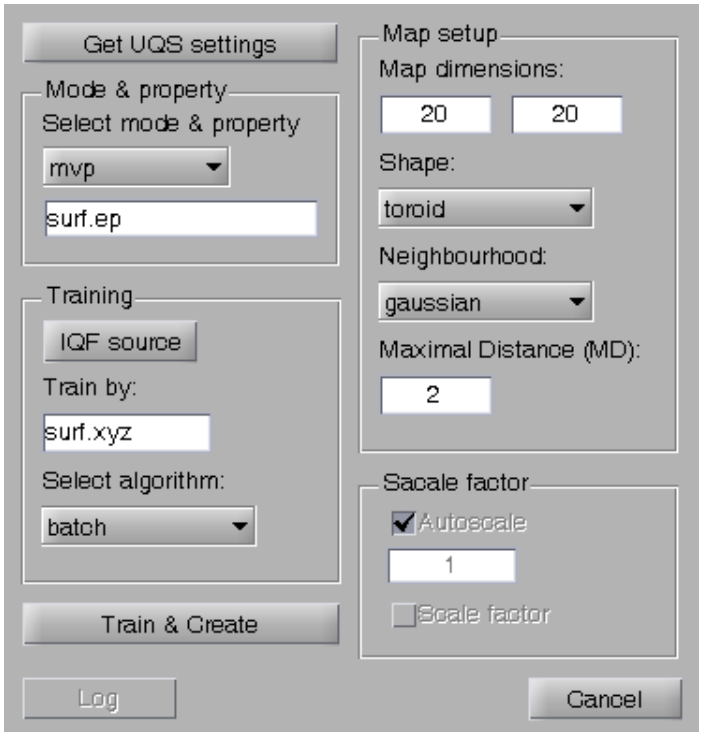
Program DDT posiada również moduł umożliwiający generowanie deskryptora SOMS-CoMSA. W rozdziale 9.2.2.2 omówiono sposób zapisu ustawień SOM-CoMSA w strukturach UQS. Moduł deskryptora SOM-CoMSA pozwala na wprowadzenie wartości pól wymienionych w tabeli 9.4 lub dane służące do ich obliczenia.

Rysunek 9.11 przedstawia okno ustawień deskryptora SOM-CoMSA. Część opcji jest podobna jak w przypadku modułu s-CoMSA (patrz rozdział 9.4.1). Ramka `Mode & property` umożliwia wybór trybu obliczania deskryptora: `mvp`, `nps`, `mvp/nps` (patrz wzory 5.3, 5.4 oraz 5.5 – strona 49). Jeżeli wybrano tryb `nps` lub `mvp/nps` istnieje możliwość podania czynnika skalującego obliczane wartości (opcja `Scale factor`). Program może również obliczyć czynnik skalujący automatycznie na podstawie gęstości stosowanej powierzchni (opcja `Autoscale`). Ramka `Mode & property` zawiera również okienko, w którym należy podać własność, w formie pola struktury IQF, używaną do obliczania wartości sektora (wymagane w przypadku trybów `mvp` oraz `mvp/nps`). W przypadku potencjału elektrostatycznego należy wprowadzić ciąg znaków: `surf.ep`. Wybrana własność musi być zgodna z polem `Train by` z ramki `Training`, które standardowo wskazuje pole przechowujące

powierzchnię. W tej samej ramce można wybrać rodzaj algorytmu trenowania mapy – lista `Select algorithm`. Przycisk `IQF source` służy do wyboru cząsteczek używanych do trenowania mapy.

Ramka `Map setup` zawiera szereg opcji dotyczących mapy. W polach `Map dimensions` należy podać rozmiar mapy. Opcja `Shape` pozwala wybrać topologię mapy a `Neighbourhood` rodzaj sąsiedztwa. Parametr MD (ang. maximal distance) może być ustawiony w okienku `Maximal Distance (MD)`.

Szczegółowe objaśnienie opcji dotyczących algorytmu trenowania i ustawień mapy można znaleźć na stronie twórców pakietu SOM Toolbox używanego przez program DDT do generowania samoorganizujących się map [130].



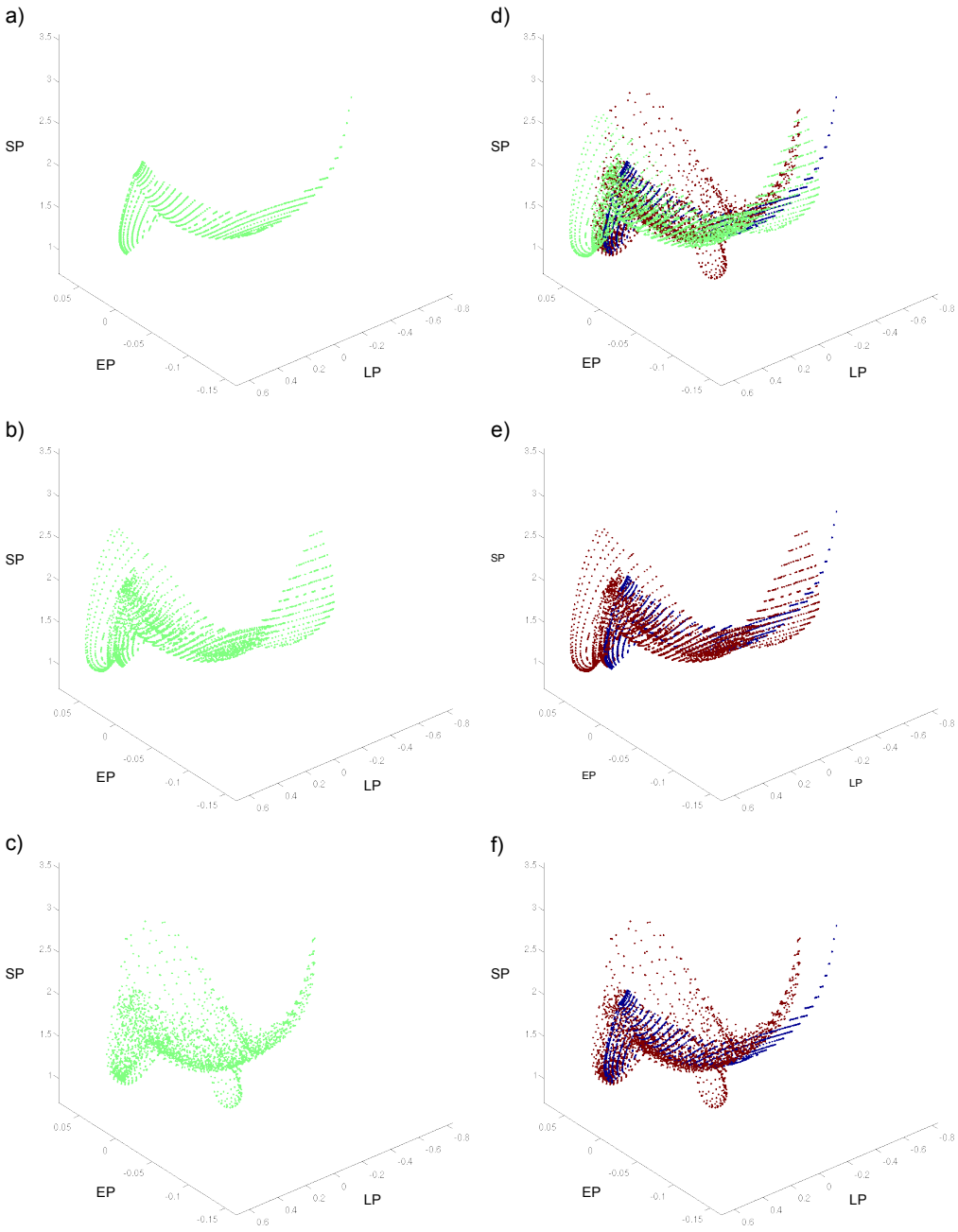
Rysunek 9.11 Okno ustawień deskryptora SOM-CoMSA programu DDT.

9.5 Hiperpowierzchnie

Wydaje się, że ciekawą koncepcją QSAR może być porównywanie hiperpowierzchni molekularnych.

Obliczone w punktach tworzących powierzchnie potencjały i własności molekularne mogą być odwzorowane na hiperpowierzchnię w przestrzeni własności. Generowanie takich hiperpowierzchni polega na przypisaniu każdemu punktowi nowych współrzędnych będących wartościami wybranych własności. Liczba nowych współrzędnych jest dowolna. Standardowo w programie DDT używane są trzy nowe współrzędne: wartość potencjału elektrostatycznego (EP), potencjału lipofilowego (LP) oraz wartość zawady sterycznej (SP). Użycie trzech współrzędnych pozwala na łatwą wizualizację hiperpowierzchni oraz i na ich dalsze przetwarzanie metodą s-CoMSA bez wprowadzania zmian w formalizmie metody. Wygenerowana powierzchnia jest przechowywana w strukturze IQF w polu **surf.ddd** a specyfikacja użytych własności w polu **surf.desc.artificial** (opis struktur IQF znajduje się w rozdziale 9.2.1).

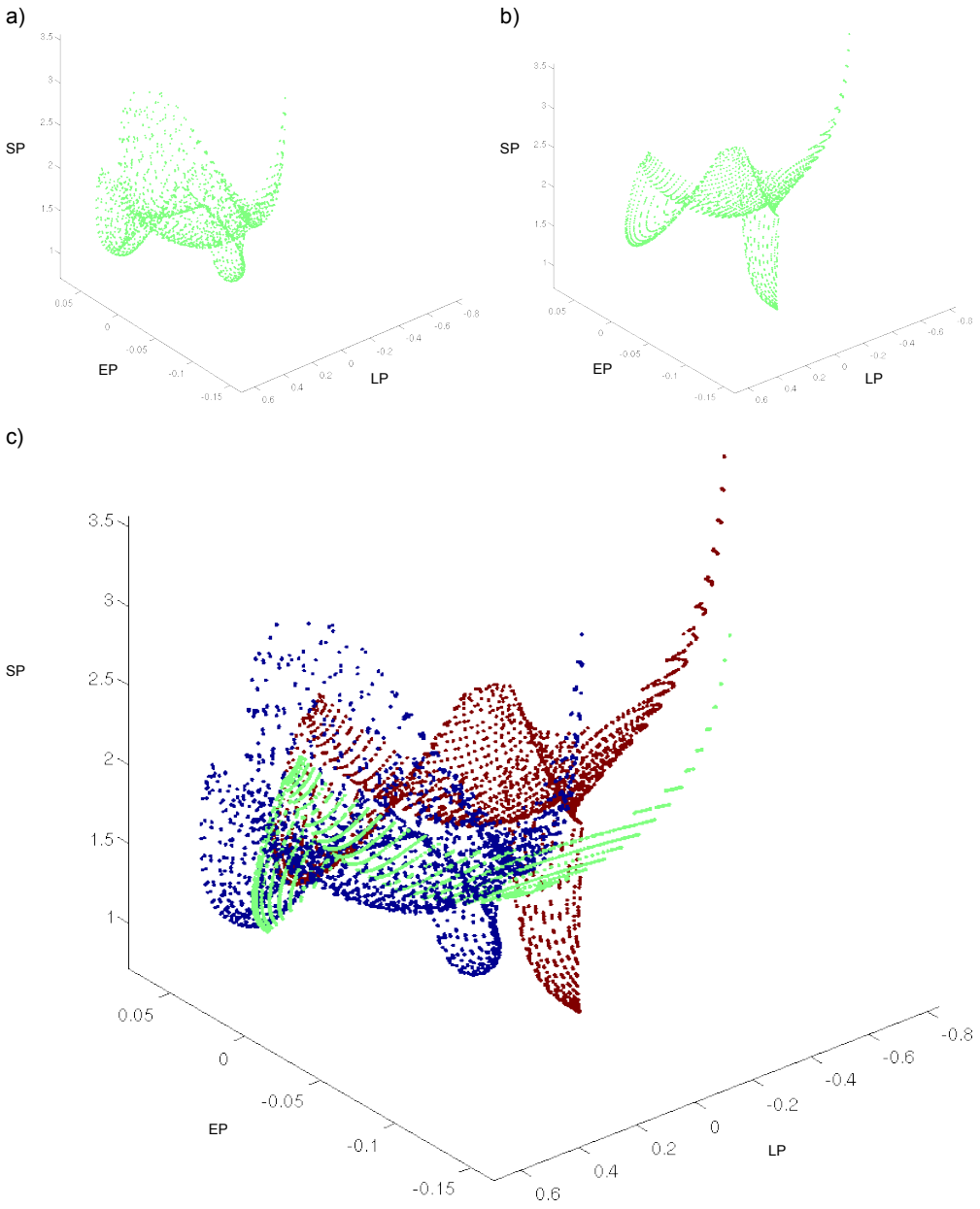
Rysunek 9.12 przedstawia hiperpowierzchnie benzenu, naftalenu oraz fenolu wygenerowane w standardowy sposób przy użyciu potencjału elektrostatycznego (EP), potencjału lipofilowego (LP) oraz wartość zawady sterycznej (SP). Niewątpliwą zaletą hiperpowierzchni jest ich samoczynne nakładanie. Hiperpowierzchnie związków o podobnych właściwościach nakładają się na siebie. Widoczne jest to na rysunku 9.12 w części d, e oraz f.



Rysunek 9.12 Hiperpowierzchnie wygenerowane w programie DDT; (a) benzen, (b) naftalen, (c) fenol, (d) benzen (niebieski), naftalen (zielony), fenol (czerwony), (e) benzen (niebieski), naftalen (czerwony), (f) benzen (niebieski), fenol (czerwony).

W przypadku związków różniących się właściwościami odpowiednie hiperpowierzchnie są na siebie nienakładalne. Rysunek 9.13 przedstawia hiperpowierzchnie hydrochinonu, chinonu oraz benzenu wygenerowane w standardowy sposób przy użyciu potencjału elektrostatycznego (EP), potencjału lipofilowego (LP) oraz

wartość zawady sterycznej (SP). Hydrochinon i chinon różnią się istotnie między sobą. Ich hiperpowierzchnie nie nakładają się na siebie. W części c rysunku 9.13 przedstawiono na jednym wykresie hiperpowierzchnie hydrochinonu, chinonu oraz dla porównania hiperpowierzchnię benzenu. Hiperpowierzchnia chinonu jest przesunięta względem hydrochinonu i benzenu. Jest to spowodowane innymi właściwościami cząsteczki chinonu.



Rysunek 9.13 Hiperpowierzchnie wygenerowane w programie DDT; (a) hydrochinon, (b) chinon, (c) benzen (zielony), hydrochinon (niebieski), chinon (czerwony).

9.6 Wstępna obróbka danych

W większości przypadków macierz obliczonych deskryptorów przed przystąpieniem do dalszej analizy powinna być poddana wstępnej obróbce. Program DDT wykonuje centrowanie lub w razie potrzeby standaryzację *in promptu* np. bezpośrednio przed wykonaniem modelowania PLS. Natomiast usuwanie zmiennych o zerowej lub niskiej wariancji jest wykonywane za pomocą odpowiedniego modułu.

Moduł wstępnej obróbki danych posiada kilka trybów pracy. Zostały one podzielone na tryby podstawowe i zaawansowane. Tryby podstawowe noszą nazwy: `only0`, `std0`, `any0`, `no0`, `nonly0`. Tryb `only0` jest domyślną metodą wstępnej obróbki – jest to tryb, w którym usuwane są zmienne posiadające wartość zero dla wszystkich obiektów.

Obróbka w trybie `std0` powoduje usunięcie zmiennych mających odchylenie standardowe równe zero. W trybie `any0` usuwane są zmienne mające wartość zero dla któregokolwiek obiektu. Tryb `no0` jest odwrotnością trybu `any0`. Powoduje on usuniecie zmiennych, które nie posiadają dla żadnego obiektu wartości zero. Jest to jedyny tryb pozostawiający zmienne zawierające wyłącznie zera (jeżeli takie występują w macierzy). Rozwinięciem trybu `no0` jest tryb `nonly0`. Powoduje on usunięcie zmiennych nieposiadających dla żadnego obiektu wartości zero oraz tych, które posiadają wyłącznie wartość zero. Jest to więc połączenie trybów `no0` i `only0`.

Pośród zaimplementowanych trybów zaawansowanej obróbki na szczególną uwagę zasługują tryby `std`, `mean`, `occur` oraz `corry`. Każdy z nich, we właściwy dla siebie sposób, oblicza na podstawie macierzy $\boldsymbol{X}$ wektor wierszowy charakteryzujący liczbowo poszczególne zmienne. Te, których charakterystyka nie spełnia zadanego warunku są usuwane z macierzy.

W trybie `std` zmienne charakteryzowane są przez obliczanie odchylenia standardowego. Tryb `mean` polega na obliczeniu średniej wartości zmiennych. Charakterystyka zmiennych w trybie `occur` jest odwrotną częstością występowania wartości zero wyrażoną liczbą z zakresu [0, 1]. Tryb `corry` charakteryzuje zmienne obliczając ich korelację z wektorem $\boldsymbol{y}$.

Dodatkowo, macierz ***X*** przed obliczaniem charakterystyki może być poddana działaniu funkcji abs( ) oraz sign( ) – wartość absolutna i znak liczby. Również już obliczona charakterystyka może być poddana działaniu tych funkcji.

Zaimplementowane tryby dają szerokie możliwości preselekcji zmiennych. Za ich pomocą możliwa jest również praktyczna realizacja deskryptorów typu łącznego i rozłącznego zdefiniowanych wzorami (5.7) oraz (5.8) – patrz rozdział 5.2, strona 47.

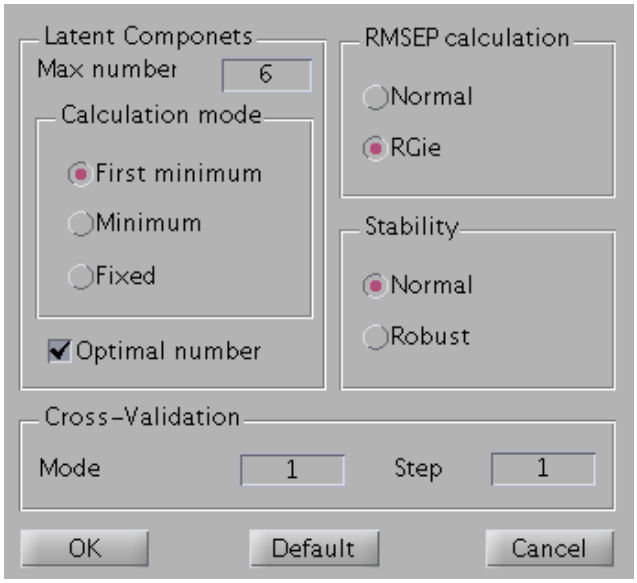
Deskryptor typu łącznego można obliczyć generując w pierwszej kolejności deskryptor dla cząsteczek referencyjnych i stosując dla nich tryb wstępnej obróbki `only0`. Wybrane kolumny zostaną zapamiętane w strukturze UQS. Po dodaniu do struktury deskryptorów pozostałych związków w dalszej analizie program DDT będzie korzystał z początkowo wybranych kolumn co w praktyce oznacza zastosowanie deskryptora typu łącznego.

W przypadku deskryptora typu rozłącznego sposób postępowania jest analogiczny do opisanego powyżej. W miejsce trybu wstępnej obróbki `only0` konieczne jest jednak zastosowanie trybu zaawansowanego `occur==0` lub `std==0` (symbol „==” oznacza „równa się”).

9.7 Model PLS

Wszystkie metody opisane w rozdziałach 3.4 oraz 3.5 (strony 30 oraz 35) zostały zaimplementowane w programie DDT. Rysunek 9.14 przedstawia okno ustawień modułu PLS. Maksymalna możliwa kompleksowość modelu może być ustawiona za pomocą pola `Max number` z ramki `Latent Componets`. Opcje `First minimum`, `Minimum` oraz `Fixed` odnoszą się do metod wyznaczania maksymalnej kompleksowości. Dwie pierwsze opcje oznaczają odpowiednio poszukiwanie pierwszego minimum *RMSEP* oraz minimum globalnego, trzecia opcja powoduje użycie maksymalnej kompleksowości jako optymalnej. Wybranie opcji `RGie` z ramki `RMSEP calculation` powoduje, że do obliczania *RMSEP* stosowany jest wzór (3.14). Zmiana metody obliczania stabilności zmiennych jest możliwa za pomocą opcji z ramki `Stability`; opcja `Normal` sprawia, że do obliczania stabilności używany jest wzór (3.15) natomiast opcja `Robust` wzór

(3.16). Stabilność obliczaną wzorem (3.17) można uzyskać posługując się linią komend. Opcja `Optimal number` powoduje, że stabilność jest liczona nie dla maksymalnej lecz dla optymalnej kompleksowości.



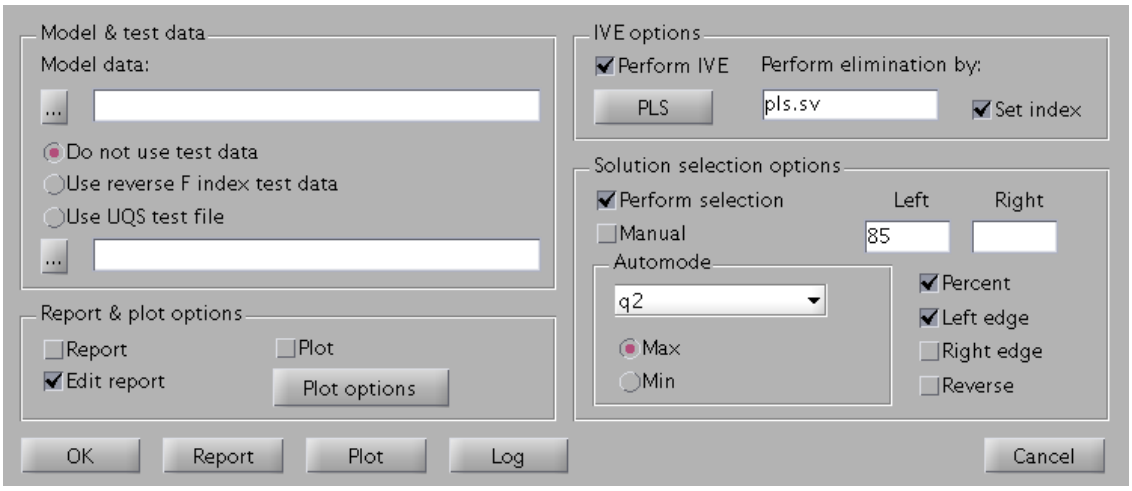
Rysunek 9.14 Okno ustawień modułu PLS programu DDT.

Zmiana rodzaju walidacji krzyżowej jest uzyskiwana przez modyfikację wartości `Mode` w ramce `Cross-Validation`. Wartość 1 oznacza walidację typu LOO. Większe wartości odpowiadają walidacji LSO. Wartość `Step` pozwala modyfikować liczbę uwzględnianych w walidacji krzyżowej podziałów obiektów. Wartość 1 oznacza uwzględnienie wszystkich podziałów, wartość 2 oznacza uwzględnienie co drugiego podziału etc.

9.8 Eliminacja zmiennych metodą IVE-PLS

Program posiada możliwość wykonania eliminacji zmiennych metodą IVE-PLS. Implementacja obejmuje algorytm opisany w rozdziale 3.6.2 (strona 37). Dodatkową opcją jest możliwość zmodyfikowania używanej stabilności za pomocą dowolnej funkcji lub zastąpienie stabilności innym parametrem obliczanym dla zmiennych przez podprogram PLS. W ogólności na przebieg procedury IVE-PLS istotny wpływ mają ustawienia opcji PLS.

Rysunek 9.15 przedstawia okno ustawień modułu IVE-PLS. Procedura może być wykonana z użyciem zbioru testowego. Zbiór testowy może być wybrany za pomocą odpowiednich opcji z ramki `Model & test data`. Jeżeli zbiór testowy zostanie użyty w czasie działania procedury dla kolejnych kroków będą obliczane parametry *SDEP* oraz r_t^2 potrzebne do uzyskania odpowiednich wykresów (patrz rozdział 7.1.1, strona 81). Przycisk `PLS` daje możliwość wyboru opcji PLS (patrz rozdział 9.7). Można również wprowadzić ręczne modyfikacje używanej stabilności – okienko `Perform elimination by`.



Rysunek 9.15 Okno ustawień modułu IVE-PLS programu DDT.

Program DDT umożliwia wybór optymalnego rozwiązania IVE na różne sposoby – odpowiednie opcje znajdują się w ramce `Solution selection options`. Każdej iteracji IVE odpowiada konkretny zestaw zmiennych. Pierwszej iteracji odpowiadają wszystkie zmienne. W każdej następnej iteracji usuwana jest jedna zmienna aż do usunięcia wszystkich. Ostatniej iteracji odpowiada więc tylko jedna zmienna. Najprostszą metodą wyboru jest szukanie iteracji charakteryzującej się maksymalną wartością q_{CV}^2 lub optymalną wartością innych parametrów. Należy zaznaczyć, że wybór w oparciu o parametry zewnętrznej walidacji (r_t^2 lub *SDEP*) powoduje, że zbiór użyty do ich obliczenia przestaje być zbiorem *stricte* zewnętrznym. Standardowo wybór optymalnej iteracji powoduje włączenie do modelu wszystkich zmiennych odpowiadających danej iteracji. Program DDT daje również możliwość wyboru zmiennych przynależnych

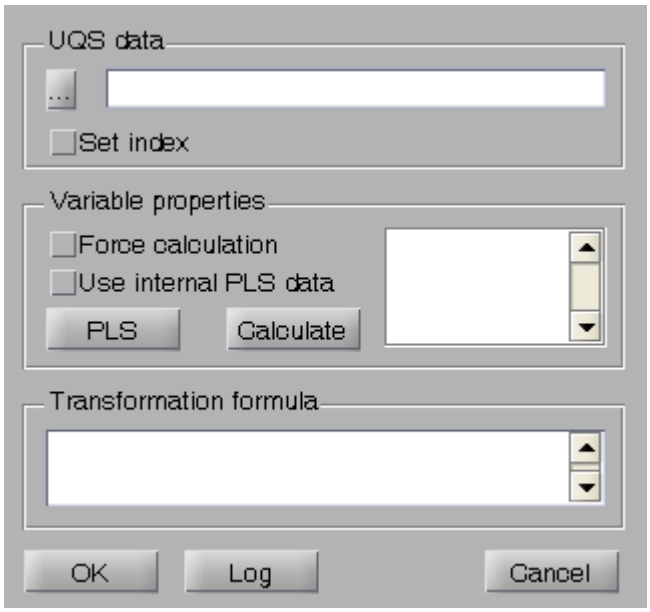
pewnemu zakresowi iteracji a nawet zmiennych, które zostały przez procedurę IVE-PLS odrzucone. Dodatkowo istnieje również ręczny tryb wyboru optymalnej iteracji – opcja `Manual`.

9.9 Identyfikacja obszarów oddziaływań specyficznych

Program DDT umożliwia wybór obszarów oddziaływań specyficznych na kilka sposobów. Identyfikacja może opierać się o metody wyboru i eliminacji zmiennych (patrz również rozdział 7.1.2, strona 85). Możliwe jest także filtrowanie własności powierzchniowych (patrz rozdział 7.2, strona 91) oraz różnych własności deskryptora QSAR. Program DDT umożliwia także łączenie wszystkich sposobów.

9.9.1 Wykorzystanie eliminacji zmiennych oraz własności deskryptora QSAR

Wynik przeprowadzonej eliminacji zmiennych jest zapisywany w plikach UQS podobnie jak wyniki modelowania metodą PLS. Rysunek 9.16 przedstawia okno ustawień identyfikacji obszarów oddziaływań specyficznych. Zapisane dane są dostępne w postaci pól widocznych okienku w prawej części ramki `Variable properties`. Pola te mogą być użyte do uzyskania zestawu zmiennych wskazujących obszary oddziaływań specyficznych. Uzyskanie tych obszarów sprowadza się do podania formuły matematycznej w oknie `Transformation formula`. Wynik działania formuły transformującej musi być wektorem o długości równiej liczbie zmiennych użytego deskryptora. Składnia formuły jest identyczna ze składnią Matlab. Wynik obliczeń jest zapamiętywany w odpowiednim polu struktury UQS, które może być powtórnie wykorzystane do wizualizacji obszarów oddziaływań specyficznych.



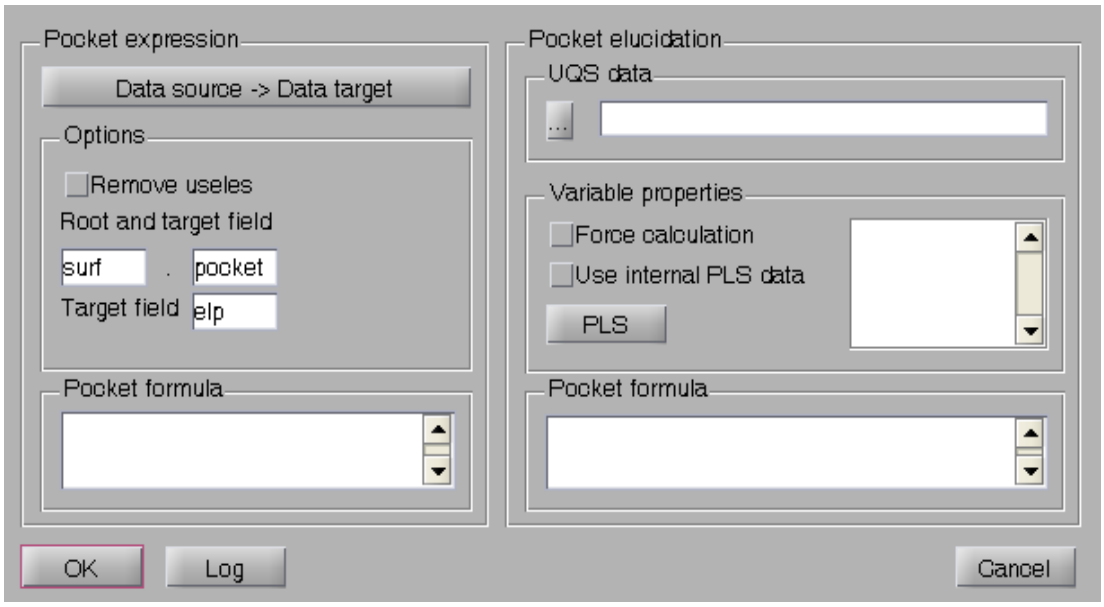
Rysunek 9.16 Okno ustawień identyfikacji obszarów oddziaływań specyficznych.

9.9.2 Filtrowanie własności cząsteczkowych

Moduł umożliwiający filtrowanie własności cząsteczkowych jest wykorzystywany zarówno do identyfikacji obszarów oddziaływań specyficznych oraz pośrednio do wizualizacji zidentyfikowanych obszarów. Umożliwia on bowiem zapis wyników identyfikacji w postaci plików IQF.

Rysunek 9.17 przedstawia okno ustawień modułu. Składa się ono z dwóch głównych części. Część po prawej stronie umożliwia wykonanie obliczeń opisanych w rozdziale powyżej. Część po lewej stronie umożliwia natomiast filtrowanie własności cząsteczkowych. Zasadniczo zasada uzyskania obszarów oddziaływań specyficznych za pomocą filtrowania własności cząsteczkowych jest analogiczna do opisanej w rozdziale powyżej. Wynik działania formuły transformującej również musi być wektorem lecz jego długość musi w tym przypadku być równa długości wektorów opisujących użyte do filtrowania własności cząsteczkowe takich jak np. potencjał elektrostatyczny, lipofilowy etc.

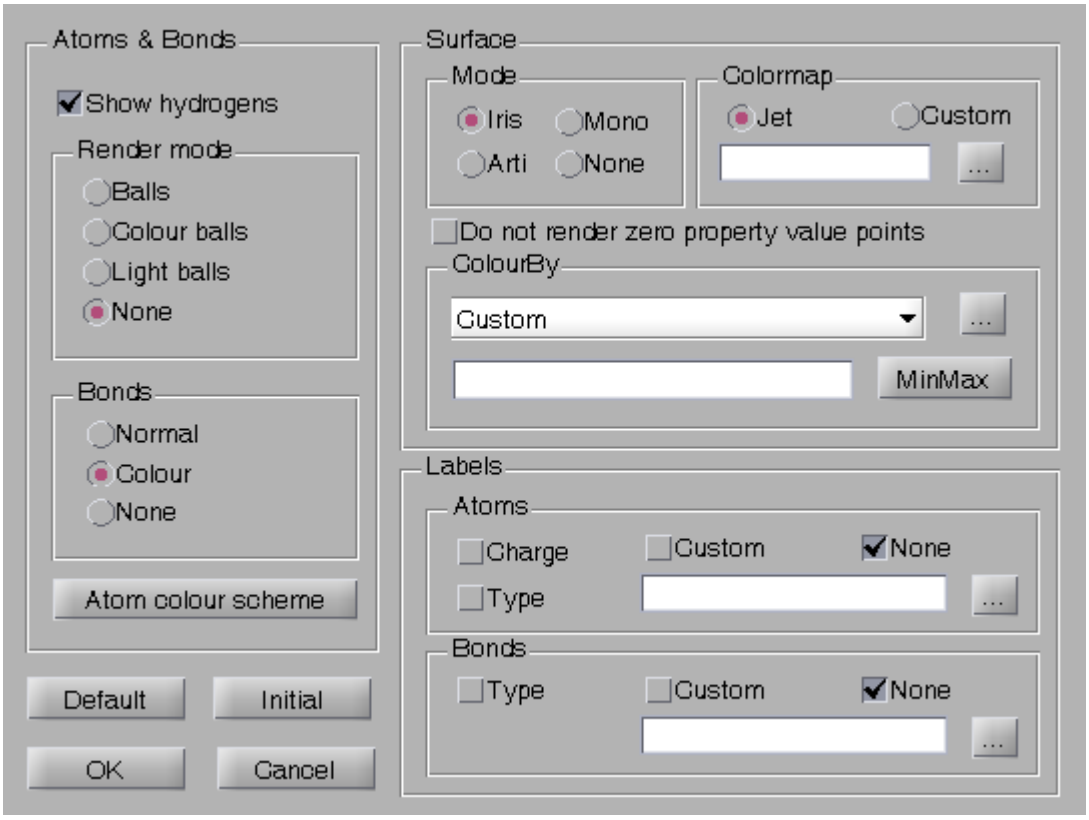
Domyślnie wynik działania formuły transformującej z prawej strony okna jest zapisywany w polu struktury IQF `pocket` (ramka `Options`, opcja `Root and target field`). Pole to może być wykorzystane w formule transformującej używanej do filtrowania własności cząsteczkowych. Wynik działania tej formuły jest domyślnie zapisywany w polu `elp` struktury IQF.



Rysunek 9.17 Okno ustawień filtrowania własności cząsteczkowych.

9.10 Wizualizacja molekuł

Podstawowe opcje wizualizacji molekuł są dostępne bezpośrednio z menu głównego. Dostęp do szczegółowych ustawień jest możliwy poprzez moduł `Molecular system manager`. Rysunek 9.18 przedstawia okno ustawień wizualizacji molekuł. Poza podstawowymi opcjami dotyczącymi sposobu wyświetlania atomów, wiązań oraz etykiet możliwe jest szczegółowe wybranie ustawień opcji dotyczących wyświetlania powierzchni cząsteczkowych a więc również sposobu wyświetlania zidentyfikowanych obszarów oddziaływań specyficznych.



Rysunek 9.18 Okno ustawień filtrowania własności cząsteczkowych.

Tryb wyświetlania `Iris` powoduje, że wskazana w ramce `ColourBy` własność jest używana do kolorowania powierzchni cząsteczkowych (własności wprowadzane w ramce `ColourBy` mogą być dodatkowo poddane działaniu dowolnych funkcji pakietu Matlab zgodnie z obowiązującą składnią). Tryb `Mono` umożliwia kolorowanie powierzchni zgodnie z czterokolorowym (również przy użyciu większej liczby kolorów) schematem opisanym w rozdziale 7.2, tabela 7.1 (strona 95). W tym celu w ramce `ColourBy` należy wprowadzić odpowiednią liczbę własności oddzielonych od siebie średnikiem.

Opcja `Do not render zero propery value points` jest bardzo użyteczna. Jej działanie polega na pomijaniu tych fragmentów powierzchni, dla których własność wprowadzona w ramce `ColourBy` wynosi zero.

9.11 Rozwój programu DDT

Program Drug Design Toolbox został napisany w sposób umożliwiający łatwą rozbudowę o nowe moduły. Obecnie zaimplementowane procedury obliczeniowe również mogą być w prosty sposób udoskonalane. Dalszy rozwój pakietu może na przykład obejmować:

- implementację wydajniejszych algorytmów PLS,
- rozbudowę modułu wyboru zmiennych o nowe metody,
- dodanie nowych modułów umożliwiających detekcję obiektów odległych oraz wybór reprezentatywnych obiektów zbiorów,
- zwiększenie liczby obsługiwanych formatów plików,
- rozbudowę modułu obsługi baz danych QDB,
- rozwój stosowanych formatów plików (IQF oraz UQS), etc.

Dodatkowo program DDT zostanie udostępniony na stosownej licencji od pobrania ze strony internetowej. Dostęp do programu obecnie jest możliwy poprzez kontakt z autorem.

10 Podsumowanie

- Opracowano nową metodę modelowania zależności 3D-QSAR s-CoMSA,
- Stwierdzono, że w wyniku optymalizacji metody s-CoMSA technika ta pozwala na otrzymywanie stabilnych modeli 3D-QSAR,
- Opracowano techniki jakościowej wizualizacji takich modeli; ich analiza pozwala na ujawnienie molekularnych uwarunkowań aktywności szeregów badanych bioefektorów,
- Zastosowano metodę s-CoMSA do modelowania zależności 3D-QSAR następujących szeregów związków organicznych:
  - steroidów o powinowactwie do globuliny wiążącej kortykosteroidy,
  - inhibitorów odwrotnej transkryptazy HIV – pochodnych 1[2-(hydroksyetoksy)metylo]-6(fenylotio)tyminy – HEPT,
  - pochodnych α-asaronu o działaniu hipolipidemicznym,
  - benzofuranowych inhibitorów N-mirystytransferazy,
  - barwników heterocyklicznych o powinowactwie do celulozy,
  - inhibitorów reduktazy kwasu foliowego – pochodnych 2,4-diamino-5-benzylpirymidyny.
- Opracowano metody ilościowej wizualizacji obszarów oddziaływań specyficznych,
- Zaimplementowano opracowaną metodę w środowisku programowania Matlab

11 Bibliografia

1 Crum-Brown, A.; Fraser, T.R.; On the Connection between Chemical Constitution and Physiological Action. Part I. - On the Physiological Action of the Salts of the Ammonium Bases, derived from Strychnia, Brucia, Thebaia, Codeia, Morphia, and Nicotia, *Trans. Roy. Soc. Edinburgh*, 25, **1869**, 151-203

2 Crum-Brown, A.; Fraser, T.R.; On the Connection between Chemical Constitution and Physiological Action. Part II. - On the Physiological Action of the Ammonium Bases derived from Atropia and Conia, *Trans. Roy. Soc. Edinburgh*, 25, **1869**, 693-739

3 Richet, M.C.; Note sur le Rapport Entre la Toxicite et les. Propriretes Physiques des Corps, *Compt. Rend. Soc. Biol. (Paris)*, 45, **1893**, 775-776

4 Meyer, H.H.; Zur theorie de alkoholnarkose. I. Mitt. Welche eigenschaft der anästhetika bedingt ihre narkotische wirkung?, *Arch. Exp. Path. Pharmakol.*, 42, **1899**, 109-118

5 Overton, E.; (Fisher, G.); Studien über die Narkose, *Jena*, **1901**

6 Hansch, C.; Fujita, T.; p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure, *J. Am. Chem. Soc.*, 86, **1964**, 1616 - 1626

7 Free, S.M.; Wilson, J.W.; A Mathematical Contribution to Structure-Activity Studies, *J. Med. Chem.*, 7, **1964**, 395 - 399

8 Polanski, J.; Wybrane problemy projektowania substancji biologicznie aktywnych, *Wiad. Chem.*, 53, **1999**, 1-16

9 Wise, M.; Cramer, R.D.; Smith, D.; Exman, I.; Progress in three-dimensional drug design: Theuse of real-time colour graphics and computer postulation of bioactive molecules in DYLOMMS, In: Quantitative Approaches to Drug Design, ed. Deardon, J.C., *Elsevier: Amsterdam*, **1983**

10 Cramer, III, R.D.; Patterson, D.E.; Bunce, J.D.; Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins, *J. Am. Chem. Soc.*, 110, **1988**, 5959-5967

11 Kubinyi, H.; QSAR and 3D QSAR in drug design. Part I: methodology, *Drug Discovery Today*, 2, **1997**, 457-467

12 Kubinyi, H.; QSAR and 3D QSAR in drug design. Part II: applications and problems, *Drug Discovery Today*, 2, **1997**, 538-546

13 Sybyl Computational Informatics Software for Molecular Modelers, http://www.tripos.com/

14 Wermuth, C.-G.; Langer, T.; Pharmacofore Idnetification, In: 3D QSAR in Drug design. Theory, Methods and Applications, ed. Kubinyi, H., *Escom, Leiden*, **1993**

15 Doweyko, A.; 3D-QSAR illusions, *J. Comput.-Aided Mol. Des.*, 18, **2004**, 587-596

16 Wold, S.; Sjostrom, M.; Eriksson, L.; PLS-regression: a basic tool of chemometric, *Chemom. Intell. Lab. Syst.*, 58, **2001**, 109-130

17 Klebe, G.; Abraham, U.; Mietzner, T.; Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity, *J. Med. Chem.*, 37, **1994**, 4130 - 4146

18 Klebe, G.; Comparative Molecular Similarity Indices Analysis: CoMSIA, *Perspect. Drug. Discov. Des.*, 12/13/14, **1998**, 87–104

19 Cramer, III, R.D.; DePriest, S.A.; Petterson, D.E.; Hetcht, P.; The developing practise of comparative molecular field analysis, In: 3D QSAR in Drug Design, Kubinyi, H., *Escom, Leiden*, **1993**

20 Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G.; Self-Organizing Molecular Field Analysis: A Tool for Structure-Activity Studies, *J. Med. Chem.*, 42, **1999**, 573-583

21 Polanski, J.; Walczak, B.; The comparative molecular surface analysis (CoMSA): a novel tool for molecular design, *Comput. Chem.*, 24, **2000**, 615-625

22 Polanski, J.; Gieleciak, R.; Bak, A.; The Comparative Molecular Surface Analysis (COMSA) - A Nongrid 3D QSAR Method by a Coupled Neural Network and PLS System: Predicting pKa Values of Benzoic and Alkanoic Acids, *J. Chem. Inf. Comput. Sci.*, 42, **2002**, 184-191

23 Polanski, J.; The non-grid technique for modeling 3D QSAR using self-organizing neural networks (SOM) and PLS analysis: Application to steroids and colchicinoids, *SAR QSAR Env. Res.*, 11, **2000**, 245-261

24 Kohonen, T.; Self-Organizing Maps, *Springer, Berlin, Heidelberg, New York*, **1995, 1997, 2001**

25 Kohonen, T.; Sefl-Organization and Associative Memory, *Springer, Berlin*, **1990**

26 Zupan, J.; Gasteiger, J.; Neural Networks in Chemistry and Drug Design, *Wiley-Vch, Verlag, Winheim*, **1999**

27 Polanski, J.; Gieleciak, R.; Magdziarz, T.; Bak, A.; Grid Formalism for the Comparative Molecular Surface Analysis: Application to the CoMFA Benchmark Steroids, Azo Dyes, and HEPT Derivatives, *J. Chem. Inf. Comput. Sci.*, 44, **2004**, 1423-1435

28 Magdziarz, T.; Lozowicka, B.; Gieleciak, R.; Bak, A.; Polanski, J.; Chilmonczyk, Z.; 3D QSAR study of hypolipidemic asarones by comparative molecular surface analysis, *Bioorganic & Medicinal Chemistry*, 14, **2005**, 1630–1643

29 Gieleciak, R.; Magdziarz, T.; Bak, A.; Polanski, J.; Modeling Robust QSAR. 1. Coding Molecules in 3D-QSAR - from a Point to SurfaceSectors and Molecular Volumes, *J. Chem. Inf. Model.*, 45, **2005**, 1447-1455

30 Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T.; Self-organizing Neural Networks for Modeling Robust 3D and 4D QSAR: Application to Dihydrofolate Reductase Inhibitors, *Molecules*, 9, **2004**, 1148-1159

31 Polanski, J.; Gieleciak, R.; The Comparative Molecular Surface Analysis (CoMSA) with Modified Uniformative Variable Elimination-PLS (UVE-PLS) Method: Application to the Steroids Binding the Aromatase Enzyme, *J. Chem. Inf. Comput. Sci.*, 43, **2003**, 656-666

32 Polanski, J.; Gieleciak, R.; Wyszomirski, M.; Mapping dye pharmacophores by the comparative molecular surface analysis (CoMSA): application to heterocyclic monoazo dyes, *Dyes Pigm.*, 62, **2004**, 63-78

33 Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T.; Modeling robust QSAR, *J. Chem. Inf. Model.*, 46, **2006**, 2310-2318

34 Hasegawa, K.; Matsuoka, S.; Arakawa, M; Funatsu, K.; New molecular surface-based 3D-QSAR method using Kohonen neural network and 3-Way PLS, *Comput. & Chem.*, 26, **2002**, 583-589

35 Todeschini, R.; Lasagni, M.; Marengo, E.; New Molecular Descriptorsfor 2D and 3D structures, Theory, *J. Chemometrics*, 8, **1994**, 263-272

36 Todeschini, R.; Gramatica, P.; Provenzani, R.; Marengo, E.; Weighted Holidtic Invariant Molecular descriptors, Part 2, Theory development and applications on modeling physico-chemical properties of PolyAromatic Hydrocarbons, *Chemom. Intell. Lab. Systems*, 27, **1995**, 221-229

37 Todeschini, R.; Bettiol, C.; Giurin, G.; Gramatica, P.; Miana, P.; Argese, E.; Modeling and prdisctions by using WHIM descriptors in QSAR studies: submitochondrial particles (SMP) as toxicity biosensors of chlorophenols, *Chemosphere*, 33, **1996**, 71-79

38 Todeschini, R.; Moro, G.; Boggia, R.; Bonati, L.; Cosentino, U.; Lasagni, M.; Pitea, D.; Modeling and predictions of molecular properties. Theory of grid-weighted holistic invariant molecular (G-WHIM) descriptors, *Chemom. Intell. Lab. Systems*, 36, **1997**, 65-73

39 Bravi, G.; Gancia, E.; Mascagni, P,; Pegna, M.; Todeschini, R.; Zaliani, A.; MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties:A comparative 3D QSAR study in a series of steroids, *J. Comput. Aided Mol. Des.*, 11, **1997**, 79-92

40 Gohlke H.; Klebe G.; DrugScore Meets CoMFA: Adaptation of Fields for Molecular Comparision (AFMoC) or How to Tailor Knowledge-Based Pair-Potentials to a Particular Protein, *J. Med. Chem.*, 45, **2002**, 4153-4170

41 Ivanciuc, O.; Ivanciuc, T.; Cabrol-Bass, D.; 3D quantitative structure activity relationships with CoRSA. Comparative receptor surface analysis. Application to calcium channel agonists, *Analusis*, 28, **2000**, 637-642

42 Wyvill, G.; McPheeters, C.; Wyvill, B.; Data structures for soft objects, *Visual Comput.*, 2, **1986**, 227-234

43 Hahn, M.; Receptor Surface Models. 1. Definition and Construction, *J. Med. Chem.*, 38, **1995**, 2080-2090

44 Hahn, M.; Rogers, D.; Receptor Surface Models. 2. Application to Quantitative Structure-Activity Relationships Studies, *J. Med. Chem.*, 38, **1995**, 2091-2102

45 Hahn, M.; Rogers, D.; Receptor surface models, *Perspect. Drug Discov. Des.*, 12/13/14, **1998**, 117-133

46 Hopfinger, A.J.; Nakata, Y.; Max, N.; QSAR of anthracycline antitumor activity and cardiac toxicity based upon intercorelation calculations, In: Intermolecular forces, ed. Pullman, B., *Reidel, Dordrecht, The Netherlands*, **1981**

47 Vedani, A.; Briem, H.; Dobler, M.; Dollinger, H.; McMasters, D.R.; Multiple conformation and protonation-state representation in 4D-QSAR: The neurokinin-1 receptor system, *J. Med. Chem.*, 43, **2000**, 4416–4427

48 Hopfinger, A.J.; Wang, S.; Tokarski, J.S.; Jin, B.; Albuquerque, M.; Madhav, P.J.; Duraiswami, C.; Construction of 3D QSAR models using the 4D QSAR analysis formalism, *J. Am. Chem. Soc.*, 119, **1997**, 10509-10524

49 Pan, D.; Tseng, Y.; Hopfinger, A.J.; Quantitative Structure-Based Design: Formalism and Application of Receptor-Dependent RD-4D-QSAR Analysis to a Set of Glucose Analogue Inhibitors of Glycogen Phosphorylase, *J. Chem. Inf. Comput. Sci.*, 43, **2003**, 1591-1607

50 Santos-Filho, O.A.; Hopfinger, A.J.; The 4D-QSAR Paradigm: Application to a Novel Set of Non-peptidic HIV Protease Inhibitors, *QSAR*, 21, **2002**, 369-381

51 Pan, D.; Liu, J.; Senese, C.; Hopfinger, A. J.; Tseng, Y.; Characterization of a Ligand-Receptor Binding Event Using Receptor-Dependent Four-Dimensional Quantitative Structure-Activity Relationship Analysis, *J. Med. Chem.*, 47, **2004**, 3075-3088

52 Vedani, A.; Dobler, M.; Multidimensional QSAR: Moving from three- to five-dimensional concepts, *Quant. Struct.-Act. Relat.*, 21, **2002**, 382–390

53 Vedani, A.; Dobler, M.; 5D-QSAR: The Key for Simulating Induced Fit?, *J. Med. Chem*, 45, **2002**, 2139-2149

54 User and Reference Manual Quasar 4.0, http://www.biograf.ch/

55 User and Reference Manual Quasar 5.2, http://www.biograf.ch/

56 Vedani, A.; Dobler, M.; Lill, M.A.; In silico prediction of harmful effects triggered by drugs and chemicals, *Toxicol. Appl. Pharmacol.*, 207, **2005**, S398-S407

57 Lill, M.A.; Dobler, M.; Vedani, A.; Multi-Dimensional QSAR in Drug Discovery: Probing Ligand Alignment and Induced Fit - Application to GPCRs and Nuclear Receptors, *Curr. Comput. Aided. Drug. Des.*, 1, **2005**, 307-324

58 Vedani, A.; Dobler, M.; Zbinden, P.; Quasi-atomistic receptor surface models: A bridge between 3-D QSAR and receptor modeling, *J. Am. Chem. Soc.*, 120, **1998**, 4471–4477

59 Receptor As Poly Tier Object Representation, http://www.biograf.ch/

60 Lill, M.A.; Vedani, A.; Dobler, M.; Raptor - combining dual-shell representation, induced-fit simulation and hydrophobicity scoring in receptor modeling: Application towards the simulation of structurally diverse ligand sets, *J. Med. Chem.*, 47, **2004**, 6174–6186

61 Lill, M.A.; Winiger, F.; Vedani, A.; Ernst, B.; Impact of induced fit on ligand binding to the androgen receptor: A multidimensional QSAR study to predict endocrine-disrupting effects of environmental chemicals, *J. Med. Chem.*, 48, **2005**, 5666–5674

62 Lill, M.A.; Dobler, M.; Vedani, A.; Prediction of small-molecule binding to Cytochrome P450 3A4: Flexible docking combined with multidimensional QSAR, *ChemMedChem*, 1, **2006**, 73-81

63 Vedani, A.; Dobler, M.; Dollinger, H.; Hasselbach, K-M.; Birke, F.; Lill, M.A.; Novel ligands for the chemokine receptor-3 (CCR3): A receptor-modeling study based on 5D-QSAR, *J. Med. Chem.*, 48, **2005**, 1515–1527

64 Vedani, A.; Dobler, M.; Lill, M.A.; Combinig Protein Modeling and 6D-QSAR. Simulating the Binding of Strucurally Diverse Ligands to the Estrogen Receptor, *J. Med. Chem.*, 48, **2005**, 3700-3703

65 Goldberg, D.E.; Genetic algorithm in search , optimization, and machine learning, *Addison-Wesley, New York*, **1989**

66 Leardi, R.; Bogia, R.; Terrile, M.; Genetic algorithms as a strategy for feature selection, *J. Chemom.*, 6, **1992**, 267-281

67 Rogers. D.; Hopfinger, A.J.; Application of genetic function approximation to QSAR and QSPR, *J. Chem. Inf. Comput. Sci.*, 34, **1994**, 854-866

68 Kubinyi, H.; Variable selection in QSAR studies. I. An evolutionary algorithm, *Quant. Struct.-Act. Relat.*, 13, **1994**, 285-294

69 Kubinyi, H.; Variable selection in QSAR studies. II. A highly eficient combination of systematic search and evolution, *Quant. Struct.-Act. Relat.*, 13, **1994**, 393-401

70 Kumura, T.; Hasegawa, K.; Funatsu, K.; GA strategy for variable selection in QSAR studies: Application fo GA-based region selection for CoMFA modeling, *J. Chem. Inf. Comput. Sci.*, 38, **1998**, 276-282

71 Hasegawa, K.; GA strategy for variable selection in QSAR studies: Application of GA-based region selection to a 3D QSAR study of acetylcholinesterase inhibitors, *J. Chem. Inf. Comput. Sci.*, 39, **1999**, 112-120

72 Cho, S.J.; Tropsha, A.; Cross-Validated R2 - Guided Region Selection for CoMFA: A simple method to achieve consitent results, *J. Med. Chem.*, 38, **1995**, 1060-1066

73 Centner, V.; Massart, D.L.; de Noord, O.E.; de Jong, S.; Vandeginste, B.M.V.; Sterna, C.; Elimination of uninformative variables formultivariate calibration, *Anal. Chim. Acta*, 330, **1996**, 1-17

74 Daszykowski, M.; Stanimirova, I. ; Walczak, B.; Daeyaert, F.; de Jong, M.R.; Heeres, J.; Koymans, L.M.H.; Lewi, P.J.; Vinkers, H.M.; Janssen, P.A.; Massart, D.L.; Improving QSAR models for biological activity of HIV Reverse Transcriptase inhibitors: Aspects of outlier detection and uninformative variable elimination, *Talanta*, 68, **2005**, 54-60

75 Barnes, R.J.; Dhanoa, M.S.; Lister, S.J.; Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra, *Appl. Spectrosc.*, 43, **1989**, 772-777

76 Naes, T.; Isaksson, T.; Fearn, T.; Davies, T.; A User-Friendly Guide to Multivariate Calibration and Classification, *NIR Publications, Wiltshire*, **2002**

77 Pearson, K.; On lines an planes of closest fit to system of points in space, *Philipine Magazine*, 2, **1901**, 559-572

78 Hoteling, H.; Analysis of a complex of statistical variables into principals components, *J. Educ. Psychol.*, 24, **1933**, 417

79 Platt, D.E.; Parida, L.; Gao Y.; Floratos, A.; Rigoutsos, I.; QSAR in grossly underdetermined systems: Opportunities and issues, *IBM J. RES. & DEV.*, 45, **2001**, 533-544

80 Wold, H.; Soft Modelling by Latent Variables: The Partial Least Squares Approach, In: Perspectives in Probablity and Statistics, ed. Gani, J., *Academic Press, London*, **1975**

81 Manne, R.; Analysis of Two Partial-Least-Squares Algorithms for Multivariate Calibration, *Chemometr. Intell. Lab. Syst.*, 2, **1987**, 187-197

82 de Jong, S.; SIMPLS: an alternative approach to partial least swuares regression, *Chemometr. Intell. Lab. Syst.*, 18, **1993**, 251-263

83 Dayal, B.S.; MacGregor, J.F.; Improved PLS Algorithms, *J. Chemometr.*, 11, **1997**, 73-85

84 Gieleciak, R.; Polanski, J.; Modeling Robust QSAR. 2. Iterative Variable Elimination Schemes for CoMSA: Application for Modeling Benzoic Acid pKa Values, *J. Chem. Inf. Model.*, 47, **2007**, 547-556

85 Kubinyi, H.; Comparative Molecular Field Analysis (CoMFA), In: Encyclopedia of Computational Chemistry, ed. Johnny Gasteiger, *John Wiley & Sons, Ltd.*, **1998**

86 Testa, B.; Purcell, W.P.; A QSAR study of sulfonamide binding tocarbonic anhydrase as test of steric models, *Eur. J. Med. Chem.*, 13, **1978**, 509-514

87 Bak, A.; Polanski, J.; A 4D-QSAR study on anti-HIV HEPT analogues, *Bioorg. Med. Chem.*, 14, **2006**, 273-279

88 Ramsden, C.A.; Quantitative drug design, Vol. 4, In: Comprehensive Medicinal Chemistry, ed. Hansch, C.; Sammes, P.G.; Taylor, J.B., *Pergamon, Oxford*, **1990**

89 Klebe, G.; Structural alignment of molecules, In: 3D QSAR in Drug Design. Theory, Methods and Applications, ed. Kubinyi, H., *Escom, Leiden*, **1993**

90 Itai, A.; Tomioka, N.; Yamada, N.; Inoue, A.; Kato, Y.; Molecular superimposition for rational drug design, In: 3D QSAR in Drug Design. Theory, Methods and Applications, ed. Kubinyi, H., *Escom, Leiden*, **1993**

91 Match3D program package, available from Professor J. Gasteiger,Computer-Chemie-Centrum, University Erlangen-Nurnberg, Germany, http://www2.ccc.uni-erlangen.de/

92 Coats, E.; The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods, *Perspect. Drug Discov. Des.*, 12/13/14, **1998**, 199-213

93 Peters, R.H.; Textile chemistry. The physical chemistry of dyeing. Vol. III, *Elsevier, Amsterdam*, **1975**

94 Timofei, S.; Schmidt, W.; Kurunczi, L.; Simon, Z.; A Review of QSAR for dye affinity for cellulose fibres, *Dyes Pigm.*, 47, **2000**, 5-16

95 Frenhc, A.D.; Battista, O.A.; Cuculo, J.A.; Gray, D.G.; Kirk-Othmer Encyclopedia of Chemical Technology, *Wiley: New York*, **1993**

96 Polanski, J.; Gieleciak, R.; Wyszomirski, M.; Comparative molecular surface analysis (CoMSA) for modeling dye-fiber affinities of the azo and antraquinone dyes, *J. Chem. Inf. Comput. Sci.*, 43, **2003**, 1754-1762

97 Timofei, S.: Schmidt, W.; Kurunczi, L.; Simmon, Z.; Sallo, A.; A QSAR study of the adsorption by cellulose fibre of antraquinone vat dyes, *Dyes Pigm.*, 24, **1994**, 267-279

98 Timofei, S.; Kurunczi, L.; Schmidt, W.; Fabian, W.M.F.; Simon, Z.; Structure-affinity binding relationships by principal component regresion analysis of antraquinone dyes, *Quant. Struct. Act. Relat.*, 14, **1995**, 444-449

99 Timofei, S.; Kurunczi. L.; Schmidt, W.; Simon, Z.; Structure-affinity binding relationships of some 4-aminobenzene derivatives for cellulose fibre, *Dyes Pigm.*, 29, **1995**, 251-258

100 Timofei, S.; Kurunczi. L.; Schmidt, W.; Simon, Z.; Lipophilicity in dye-cellulose fibre binding, *Dyes Pigm.*, 32, **1996**, 25-42

101 Fabian, W.M.F. Timofei, S.; Kurunczi. L.; Comparative molecular field analysis (CoMFA), semiempirical (AM1) molecular orbital and multiconformational minimal steric difference (MTD) calculation of antraquinone dye-fibre affinities, *J. Mol. Struct. THEOCHEM*, 340, **1995**, 73-81

102 Fabian, W.M.F.; Timofei, S.; Comparative molecular field analysis (CoMFA) of dye-fibre affinities II: symmetrical bisazo dyes, *J. Mol. Struct. THEOCHEM*, 362, **1996**, 155-162

103 Oprea, T.I.; Kurunczi, L.; Timofei, S.; QSAR studies of disperse azodyes towards the negation of the pharmacophore theory of dye - fibre interaction?, *Dyes Pigm.*, 33, **1997**, 41-64

104 Funar-Timofei, S.; Schuurmann, G.; Comparative molecular field analysis (CoMFA) of anionic azo dye-fiber affinities I: gas-phase molecular orbital descriptors, *J. Chem. Inf. Comput. Sci.*, 42, **2002**, 788-795

105 Funar-Timofei, S.; Schruumann, G.; Comparative Molecular Field Analysis (CoMFA) of Anionic Azo Dye-Fiber Affinities I: Gas-Phase Molecular Orbital Descriptors, *J. Chem. Inf. Comput. Sci.*, 42, **2002**, 788-795

106 Timofei, S.; Fabian, W.M.F.; Comparative molecular field analysis(CoMFA) of heterocyclic monoazo dye-fiber affinities, *J. Chem. Inf.Comput. Sci.*, 38, **1998**, 1218-1222

107 Jalali-Heravi, M.; Parastar, F.; Use of artificial neural networks in a QSAR study of anti-HIV activity for a large group of HEPTderivatives, *J. Chem. Inf. Comput. Sci.*, 40, **2000**, 147-154

108 Kireev, D.B.; Chretien, J.R.; Grierson, D.S.; Monneret, C.; A 3DQSAR study of a series of HEPT analogues: the influence ofconformational mobility on HIV-1 reverse transctriptase inhibition, *J. Med. Chem.*, 40, **1997**, 4257-4264

109 Hannongbua, S.; Nivesanond, K.; Lawtrakul, L.; Pungpo, P.; Wolschann,P.; 3D-Quantitative structure-activity relationships of HEPT derivativesas HIV-1 reverse trascriptase inhibitors, based on ab initio calculations, *J. Chem. Inf. Comput. Sci.*, 41, **2001**, 848-855

110 Luco, J.M.; Ferretii, F.H.; QSAR based on multiple linear regressionand PLS methods for the anti-HIV activity of a large group of HEPT derivatives, *J. Chem. Inf. Comput. Sci.*, 37, **1997**, 392-401

111 Douali, L.; Villemin, D.; Charquaoui, D.; Comparative QSAR based on neural networks for the anti-HIV activity of HEPT derivatives, *Curr. Pharm. Des.*, 9, **2003**, 1817-1826

112 Mager, P.P.; Hybrid canonical-correlation neural-network approachapplied to nonnucleoside HIV-1 reverse transcriptase inhibitors (HEPT derivatives), *Curr. Med. Chem.*, 10, **2003**, 1643-1659

113 Douali, L.; Villemin, D.; Charquaoui, D.; Neural networks: accurate non linear QSAR model for HEPT derivatives, *J. Chem. Inf. Comput.Sci.*, 43, **2003**, 1200-1207

114 Gayen, S.; Debnath, B.; Samanta, S.; Jha, T.; QSAR study on some anti-HIV HEPT analogues using physicochemical and topological parameters, *Bioorg. Med. Chem.*, 12, **2004**, 1493-1503

115 Kennard, R.W.; Stone, L.A.; Computer aided design of experiments, *Technometrics*, 11, **1969**, 137-148

116 Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel onDetection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) *JAMA*, **2001**, 285, 2486

117 Levine, G.; Keaney, J.; Vita, J.; Cholesterol Reduction in Cardiovascular Disease - Clinical Benefits and Possible Mechanisms, *N. Engl. J. Med.*, 332, **1995**, 512-521

118 Szapary, P.O.; Rader, D.J.; The triglyceride–high-density lipoprotein axis: An important target of therapy?, *Am. Heart. J.*, 148, **2004**, 211-221

119 Chamorro, G.; Garduno, L.; Sanchez, A.; Labarrios, F.;Salazar, M.; Martinez, E.; Diaz, F.; Tamariz, J.; Hypolipidaemic activity of dimethoxy unconjugated propenyl side-chain analogs of α-asarone in mice, *Drug Dev.Res.*, 43, **1998**, 105-108

120 Labarrios, F.; Garduno, L.; Vidal, M.; Garcia, R.; Salazar, M.; Martinez, E.; Diaz, F.; Chamorro, G.; Tamariz, J.; Synthesis and Hypolipidaemic Evaluation of a Series of α-Asarone Analogues Related to Clofibrate in Mice, *J. Pharm. Pharmacol*, 51, **1999**, 1-7

121 Johnson, D.R.; Bhatnagar, R.S.; Knoll, L.; Gordon, J.I.; Genetic and Biochemical Studies of Protein N-Myristoylation, *Annu. Rev. Biochem.*, 63, **1994**, 869-914

122 Weinberg, R.A.; McWherter, C.A.; Freeman, S.K.; Wood, D.C.; Gordon, J.I.; Lee, S.C.; Genetic studies reveal that myristoylCoA:protein N-myristoyltransferase is an essential enzyme in Candida albicans , *Mol. Microbiol.*, 16, **1995**, 241-250

123 Lodge, J.K.; Jackson-Machelski, E.; Toffaletti, D.L.; Perfect, D.L.; Gordon, J.I.; Targeted Gene Replacement Demonstrates that Myristoyl-CoA:Protein N-Myristoyltransferase is Essential for Viability of Cryptococcus neoformans , *Proc. Natl. Acad. Sci.*, 91, **1994**, 12008-12012

124 Hasegawa, K.; Morikami, K.; Shiratori, Y.; Ohtsuka, T.; Aoki, Y.; Shimma, N.; 3D-QSAR study of antifungal N-myristoyltransferase inhibitors by comparative molecular surface analysis, *Chemom. Intell. Lab. Syst.*, 69, **2003**, 51-59

125 Polanski, J.; Drug design using comparative molecular surface analysis, *Expert Opin. Drug Discov.*, 7, **2006**, 693-707

126 Golbraikh. A.; Thropsha, A.; Beware of q2!, *J. Mol. Graph. Mod.*, 20, **2002**, 269-276

127 Polanski, J.; Gieleciak. R.; Bak, A.; Probability Issues in Molecular Design: Predictive and Modeling Ability in 3D-QSAR Schemes, *Comb. Chem. High Throughput Screen.*, 7, **2004**, 793-807

128 MATLAB 5.0 program. Available from: The MathworksInc., Natick, MA, USA, http://www.mathworks.com/

129 Ghose, A.; Viswanadhan, V.; Wendoloski, J.; Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods, *J. Phys. Chem. A.*, 102, **1998**, 3762-3772

130 SOM Toolbox, Copyright (C) 2000-2005 by Esa Alhoniemi, Johan Himberg, Juha Parhankangas and Juha Vesanto, http://www.cis.hut.fi/projects/somtoolbox/

131 CORINA Generation of 3D coordinates, http://www.mol-net.com/software/corina/

132 Gasteiger, J. et al. CTX Keyword Reference Manual, University of Erlangen-Nürnberg: 1995, unpublished results

133 Audry, E.; Dubost, J.P.; Colleter, J.C.; Dallet, P.; A New Approach to Structure-activity Relations: the "Molecular Lipophilicity Potential", *J. Med. Chem.*, 21, **1986**, 71-72

Curriculum vitae – mgr. Tomasz Magdziarz

Imię i nazwisko	Tomasz Magdziarz
Adres zamieszkania	ul. Krzywa 9 34-300 Żywiec
Data i miejsce urodzenia	24 marzec 1979, Żywiec
Wykształcenie	2003 – Rozpoczęte studia doktoranckie przy Instytucie Fizyki Uniwersytetu Śląskiego. 1998 – 2003 Studia magisterskie na wydziale Mat. Fiz. Chem. Uniwersytetu Śląskiego, kierunek – chemia. 2002 Università degli Studi di Perugia Program Socrates-Erasmus 1994 – 1998 Liceum Ogólnokształcące nr 1 im. Stanisława Staszica w Jastrzębiu Zdroju.
Znajomość języków obcych	Język angielski

Dorobek naukowy – mgr Tomasz Magdziarz

Publikacje:

1. Polanski, J.; Bak, A.; Gieleciak, R.; **Magdziarz, T.**; Modeling robust QSAR, *J. Chem. Inf. Model.*, **2006**, 46, 2310-2318
2. **Magdziarz, T.**; Lozowicka, B.; Gieleciak, R.; Bak, A.; Polanski, J.; Chilmonczyk, Z.; 3D-QSAR study of hypolipidemic asarones by comparative molecular surface analysis, *Bioorg. Med. Chem.*, **2006**, 14, 1630-1643
3. Gieleciak, R.; **Magdziarz, T.**; Bak, A.; Polanski, J.; Modeling robust QSAR. 1. Coding molecules in 3D-QSAR - from a point to surface sectors and molecular volumes, *J. Chem. Inf. Model.*, **2005**, 45, 1447-1455
4. Polanski, J.; Bak, A.; Gieleciak, R.; **Magdziarz, T.**; Self-organizing neural networks for modeling robust 3D and 4D QSAR: Application to dihydrofolate reductase inhibitors, *Molecules*, **2004**, 9, 1148-1159
5. Polanski, J.; Gieleciak, R.; **Magdziarz, T.**; Bak, A.; The grid formalism for the comparative molecular surface analysis: Application to the CoMFA benchmark steroids, azo dyes and HEPT derivatives, *J. Chem. Inf. Comput. Sci.*, **2004**, 44, 1423-1435
6. Polanski, J.; Niedbala, H.; Musiol, R.; Tabak, D.; Podeszwa, B.; Gieleciak, R.; Bak, A.; Palka, A.; **Magdziarz, T.**; Analogues of the styrylquinoline and styrylquinazoline HIV-1 integrase inhibitors: design and synthetic problems, *Acta Pol Pharm.*, **2004**, 61, 3-4

Konferencje i komunikaty:

1. **Magdziarz, T.**; Robust multidimensional QSAR modeling, Gliwice Scientific Meetings 2007, Gliwice, Poland
2. **Magdziarz, T.**; Drug-receptor interactions as seen from the perspective of 3D-QSAR methods, IV Ogólnopolskie Seminarium Doktorantów Wydziałów Chemicznych “Na pograniczu biologii i chemii”, 2006, Nachod, Czech Republic
3. Bak, A.; Polanski, J.; **Magdziarz, T.**; Gieleciak, R.; Application of the self-organizing neural networks for modeling steric and electronic effects in 4D-QSAR schemes, QSAR & Molecular Modelling in Rational Design of Bioactive Molecules, QSAR 2004, Istanbul, Turkey
4. Bak, A.; **Magdziarz, T.**; Gieleciak, R.; Polanski, J.; Self-organizing neural networks (SOM) for modeling robust 3D and 4D-QSAR: application to dihydrofolate reductase inhibitors, QSAR & Molecular Modelling in Rational Design of Bioactive Molecules, QSAR 2004, Istanbul, Turkey
5. Gieleciak, R.; **Magdziarz, T.**; Bak, A.; Polanski, J.; The comparative molecular surface analysis: new formalism and application in 3D QSAR studies, QSAR & Molecular Modelling in Rational Design of Bioactive Molecules, QSAR 2004, Istanbul, Turkey
6. **Magdziarz, T.**; Designing drugs by docking ligands into proteins, Gliwice Scientific Meetings 2003, Gliwice, Poland
7. **Magdziarz, T.**; Rogóz, R.; Gieleciak, R.; Bąk, A.; Polański, J.; Modeling dye-cellulose affinities by the grid version of the Comparative Molecular Surface Analysis, European Congress of Young Chemists "YoungChem 2003" Zakopane, Poland

Załącznik – najważniejsze publikacje dotyczące wyników omawianych badań

Na kolejnych stronach znajdują się przedruki publikacji dotyczących wyników badań omawianych w niniejszej pracy.

GRID Formalism for the Comparative Molecular Surface Analysis: Application to the CoMFA Benchmark Steroids, Azo Dyes, and HEPT Derivatives

Jaroslaw Polanski,* Rafal Gieleciak, Tomasz Magdziarz, and Andrzej Bak

Department of Organic Chemistry, Institute of Chemistry, University of Silesia, PL-40-006 Katowice, Poland

Received January 21, 2004

Shape analysis is a powerful tool in chemistry and drug design, and molecular surface defines shape in the molecular scale. In the current publication we presented a novel formalism for the comparative molecular surface analysis (s-CoMSA). The method enables both quantitative modeling of 3D-QSAR and finding possible pharmacophoric sites. The method provides very predictive models for the CBG activity of the benchmark steroid series, tinctorial properties of the heterocyclic azo dyes and anti-HIV activity of the HEPT series.

INTRODUCTION

Shape is one of the fundamental categories used by the human brain for the perception and description of 3D objects. It is the particular way in which the external edges or boundaries of objects are connected with each other that determines their shapes. Molecular surfaces model shapes of molecular objects. Although molecular surfaces are only a conventional imitation of the molecular boundaries, it has been shown that such models can often explain fundamental chemical or pharmacological effects; therefore; molecular shape is an important property that determines molecular recognition and drug−receptor interactions.

On the other hand, a number of three-dimensional Quantitative Structure Activity Relationships (3D-QSAR) methods do not use direct shape descriptors. In particular, the Comparative Molecular Field Analysis (CoMFA),[1] the first technique developed for modeling and analyzing 3D-QSARs is probably the most typical example here. This method constructs a spatially uniform 3D field around a series of superimposed molecules to investigate molecular environment, therefore, masking explicit shape information by the regularity of the cubic grid used. On the contrary, other approaches, that include explicit shape information, are well-known in molecular design. Hopfinger's Molecular Shape Analysis,[2] Receptor Surface Models,[3−5] Comparative Receptor Surface Analysis (CoRSA),[6] and Compass[7] are some of the recent methods. Direct comparison of the molecular objects usually needs a special tool that normalizes individual shapes of a series of molecules. Different techniques have been suggested to achieve this. Thus, a virtual receptor represented by the van der Waals spheres is constructed around an entire set of molecules superimposed in the receptor surface model method. Shape grids are constructed in Receptor Surface Analyses, e.g. in the CoRSA method, while the Compass uses a supervised neural network to fit two nonidentical points near the molecular surfaces. Similarly, in the Comparative Molecular Surface Analysis (CoMSA)[8] the unsupervised SOM neural network[9] is applied, that compares two (unnecessarily identical) slices of the molecular surfaces of two different molecules using the ability of the SOM neuron to group the patterns (vectors) located near the template attractor. This method enabled us to model 3D-QSAR by the analysis of various surface properties, e.g. the electrostatic potential. We proved that this technique could be an efficient tool for modeling 3D-QSAR or exploring molecular diversity.[10−16] Generally, modeling and predictive ability of SOM-CoMSA, including the comparative Kohonen neural network[17] coupled with the Partial Least Squares (PLS) method,[18] outperforms this of the CoMFA.[10−14] Recently, Hasegawa has also proved the efficiency of similar SOM-CoMSA schemes in 3D-QSARs.[19−21]

Although the SOM network has certain advantages, as it can both model nonlinear systems and preserve the topology of the objects projected,[22] the indeterministic behavior of neural architecture and the need for the special software packages that realize SOM algorithm can cause some problems. In the current research we developed a nonneural, sector version of the CoMSA (s-CoMSA) based on the grid formalism similar to that of the Hopfinger's 4D-QSAR.[2] This method has been applied for modeling 3D-QSAR of the CoMFA steroid benchmark series,[8] heterocyclic azo dye series,[13,23] and HEPT HIV blocking agents.[24]

EXPERIMENTAL SECTION

Model Builders. All the experimental data, i.e., biological activities for the CoMFA steroids, HEPT analogues, and cellulose affinities for the heterocyclic azo dyes, were extracted from the refs 8, 24, and 23 and are given in Tables 1−3, respectively.

All the molecules were superimposed before the calculation of molecular surfaces. The superimposition was performed by covering:

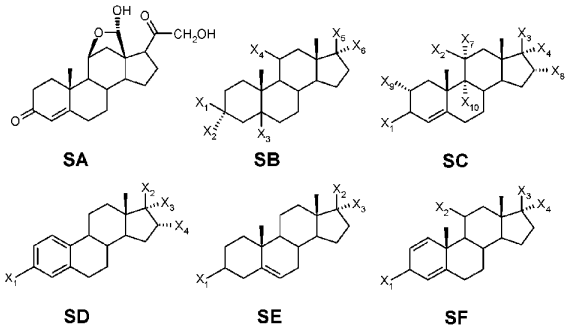
- for molecules (**s1**−**s31**) all non-hydrogen atoms of four steroid rings − molecule s6 (see Table 1),

- for molecules (**d1**−**d21**) different superimposition modes were tested, as discussed in detail further,

* Corresponding author e-mail: Polanski@us.edu.pl.

10.1021/ci049960l CCC: $27.50
Published on Web 07/08/2004

Table 1. Steroid Structures and the CBG Affinity Data[8]



no.	S	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	CBG
s1	SA											−6.279
s2	SB	OH	H	H[a]	H	OH	H					−5.000
s3	SE	OH	OH	H								−5.000
s4	SC	=O	H	=O				H	H	H	H	−5.763
s5	SB	H	OH	H[a]	H	=O						−5.613
s6	SC	=O	OH	$COCH_2OH$	H			H	H	H	H	−7.881
s7	SC	=O	OH	$COCH_2OH$	OH			H	H	H	H	−7.881
s8	SC	=O	=O	$COCH_2OH$	OH				H	H	H	−6.892
s9	SE	OH	=O									−5.000
s10	SC	=O	H	$COCH_2OH$	H			H	H	H	H	−7.653
s11	SC	=O	H	$COCH_2OH$	OH			H	H	H	H	−7.881
s12	SB	=O		H[a]	H	OH	H					−5.919
s13	SD	OH	OH	H	H							−5.000
s14	SD	OH	OH	H	OH							−5.000
s15	SD	OH	=O		H							−5.000
s16	SB	H	OH	H[b]	H	=O						−5.255
s17	SE	OH	COMe	H								−5.255
s18	SE	OH	COMe	OH								−5.000
s19	SC	=O	H	COMe	H			H	H	H	H	−7.380
s20	SC	=O	H	COMe	OH			H	H	H	H	−7.740
s21	SC	=O	H	OH	H			H	H	H	H	−6.724
s22	SF	=O	OH	$COCH_2OH$	OH							−7.512
s23	SC	=O	OH	$COCH_2OCOMe$	OH			H	H	H	H	−7.553
s24	SC	=O	=O	COMe	H				H	H	H	−6.779
s25	SC	=O	H	$COCH_2OH$	H			OH	H	H	H	−7.200
s26	SC[c]	=O	H	OH	H			H	H	H	H	−6.144
s27	SC	=O	H	COMe	OH			H	OH	H	H	−6.247
s28	SC	=O	H	COMe	H			H	Me	H	H	−7.120
s29	SC[c]	=O	H	COMe	H			H	H	H	H	−6.817
s30	SC	=O	OH	$COCH_2OH$	OH			H	H	Me	H	−7.688
s31	SC	=O	OH	$COCH_2OH$	OH			H	H	Me	F	−5.797

[a] 5-α. [b] 5-β. [c] H instead Me at the C_{10}.

- and for molecules (**h1**−**h107**) all non-hydrogen atoms of pyrimidine ring.

We used Match3D program[25] for performing this operation.

4D-QSAR Calculation. We used Hopfinger's spatial grid system[2] for coding molecules. The molecules after AM1 optimization were used as initial structures in the molecular dynamic simulation (MDs). Each 3D structure is the starting point in generating conformational ensemble profile (CEP). Molecular dynamics was performed using the Sybyl software[26] with standard Tripos force field. 2500 conformations were sampled for each analogue. Partial atomic charges were calculated using the semiempirical AM1 Hamiltonian (HYPERCHEM package[27]). The alignment of the molecules was the next step of the 4D-QSAR analysis. We aligned the molecules according to the previous rules of the CoMFA study.[1] Individual conformers are placed in the grid cell space surrounding the aligned compounds. We applied cubic grid lattice of 20 Å on each side with grid cell resolution of 1, 2, or 0.5 Å, respectively. Different types of grid cell occupancy descriptors (GCODs) were considered and calculated for the indicated atoms referred to as interaction pharmacophore elements (IPE).[2] Apart from, the GCODs used by Hopfinger,[2] we applied in our current work the absolute charge occupancy (A_q) for the chosen IPE atoms of compound c defined as

$$A_q(c,i,j,k,N) = \sum_{t=0}^{T} O_t(c,i,j,k) \times q/m \qquad (1)$$

where m means the number of the atoms of compounds, c present in the cell (i,j,k) at time t, q means the sum of partial atoms of charges present in some cell at time t, and T is the length of the time in MDs. N is the number of sampling

MDs steps. The joint (J_q) and self-charge occupacy (S_q) with the most active reference compound R defined after following equations:

$$J_q(c,i,j,k,N) = \sum_{t=0}^{T} O_t(c,i,j,k) \cap O_t(R_q,i,j,k) \times q/m \quad (2)$$

$$S_q(c,R,i,j,k,N) = \sum_{t=0}^{T} \{O_t(c,i,j,k) - [\sum_{t=0}^{T} O_t(c,i,j,k) \cap O_t(R,i,j,k)]\} \times q/m \quad (3)$$

We used the MATLAB[28] environment to program the calculation of the above-mentioned descriptors. The Partial Least Squares (PLS) method with variable elimination was used to estimate the relationship between independent variables (GCODs) and corticosteroid binding globulin (CBG) affinity.

Calculation of the Molecular Surface (s-COMSA) Descriptors based on Virtual Cubic Grid. For the calculation of shape descriptors we applied a formalism similar to Hopfinger's 4D-QSAR grid coding system using the absolute type descriptors, as given by the above-mentioned equations. However, unlike in 4D-QSAR our method compares single conformers. Thus, each 3D molecular representation is placed in its own virtual cubic grid, and the molecular surface is calculated, respectively. The electrostatic potential is calculated for the points randomly sampled on the molecular surface and a mean value of the electrostatic potential corresponding to the respective points found in each grid cell is used to describe this cell. Grid cells are unfolded into vectors and vectors describing all molecules of the series are aligned into a matrix. Grid cells that are empty for all molecules in the series analyzed are eliminated, and the resulting matrix was used for further calculations using the PLS method.

PLS Analysis. Vectors obtained were processed by the PLS analysis with a leave-one-out cross-validation procedure. The PLS procedures were programmed within the MATLAB environment (MATLAB).[28]

A PLS model was constructed for the centered data, and its complexity was estimated on the basis of the leave-one-out cross-validation procedure (CV). In the leave-one-out CV one repeats the calibration m times, each time treating the ith left-out object as the prediction object. The dependent variable for each left-out object is calculated on the basis of the model with one, two, three, etc. factors. The Root Mean Square Error of CV for the model with j factors is defined as

$$\mathrm{RMSECV}_j = \sqrt{\frac{\sum_i (\mathrm{obs}_i - \mathrm{pred}_{i,j})^2}{m}} \quad (4)$$

where obs denotes the assayed value, pred is the predicted value of the dependent variable, and i refers to the object index, which ranges from 1 to m. The model with k factors, for which RMSECV reaches a minimum, is considered as an optimal one.

We used the performance metrics that are accepted and widely used in CoMFA analyses, i.e., cross-validated q^2

$$q^2 = 1 - \frac{\sum(\mathrm{obs}_i - \mathrm{pred}_i)^2}{\sum(\mathrm{obs}_i - \mathrm{mean(obs)})^2} \quad (5)$$

where obs refers to the assayed values, pred refers to the predicted values, mean refers to the mean value of obs, and i refers to the object index, which ranges from 1 to m; and the cross-validated standard error s

$$s = \sqrt{\frac{\sum(\mathrm{obs}_i - \mathrm{pred}_i)^2}{m - k - 1}} \quad (6)$$

where m is the number of objects, and k is the number of the PLS factors in the model.

Before the PLS analysis was performed the descriptors were centered, and this operation was repeated for each cross-validation run.

The quality of external predictions was measured by the SDEP parameter

$$\mathrm{SDEP} = \sqrt{\frac{\sum_i(\mathrm{pred}_i - \mathrm{obs}_i)^2}{n}} \quad (7)$$

where pred is the predicted value, obs is the observed value, mean is the mean value, n is a number of measurements, and opt is a number of the PLS factors used in the model.

Data Elimination. To find these parts on the molecular surface that contribute to activity we used the modified procedure of the PLS with Uninformative Variable Elimination (UVE-PLS), namely the Iterative Variable Elimination PLS (IVE-PLS) procedure.[11] The UVE algorithm was originally developed by Centner et al.[29] as a possible improvement of PLS models. The purpose of the method is to reduce the number of the variables included in the final PLS model. The UVE algorithm is based on the analysis of the regression coefficients calculated by the PLS method. The PLS method allows for presenting the relation between the Y answer and the X predictors in a form of

$$Y = Xb + e$$

where b is a vector of the regression coefficients and e is the vector of the errors.

Thus, the UVE algorithm analyzes the reliability of the mean(b)/s(b) ratio (where s(b) means standard deviation of b). Then only the variables of the "relative" high mean(b)/s(b) ratio are included into the final PLS model. To estimate the cutoff level artificial random number noise is created (the level of the noise is 10^{-10} of the original variable order) and put (as additional columns) into the matrix of the original variables. The PLS analysis of such a matrix is performed and the mean(b)/s(b) parameter is analyzed for each column. The highest absolute value, abs(mean(b)/s(b)) for the noisy column, determines the cutoff level for the original variables. In the current publication we used both typical UVE-PLS[29] and its modified iterative version IVE-PLS procedure (for the detailed description see ref 11).

RESULTS AND DISCUSSION

The description of the molecular shape by spatial sectors was originally suggested by Purcell and Testa[30] and then

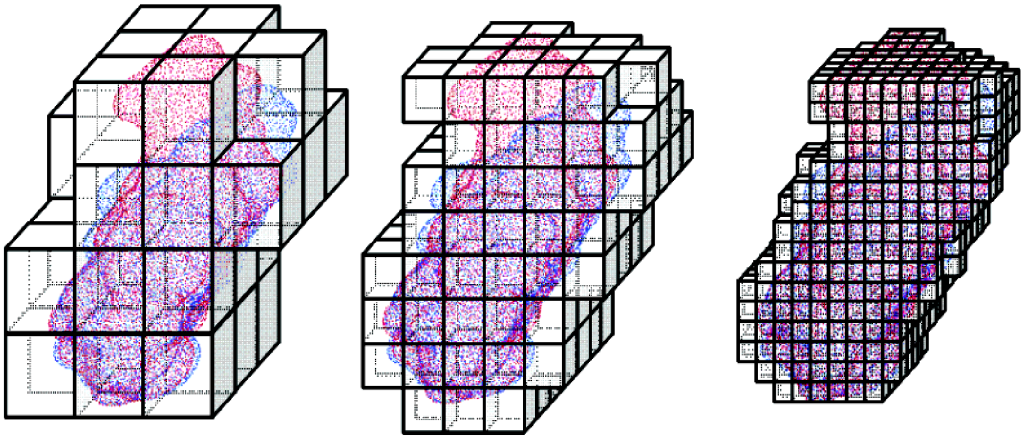


Figure 1. The molecular surface formed by a polycube grid of the different resolution.

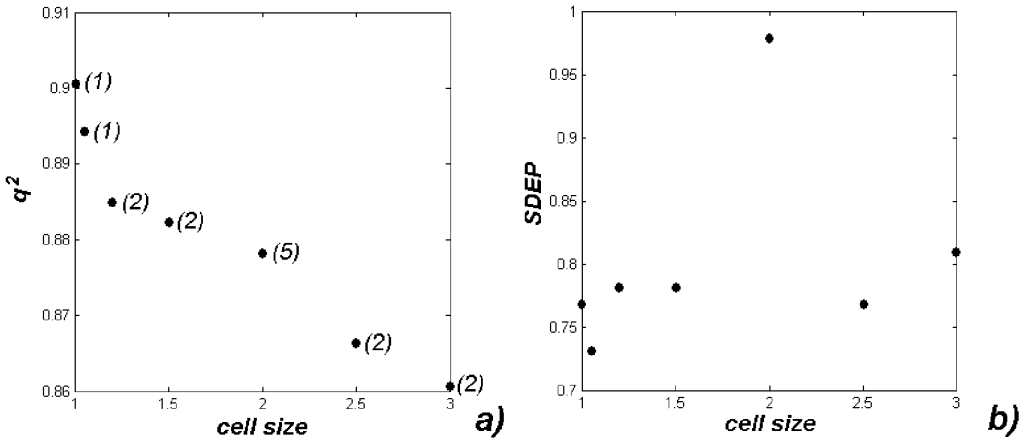


Figure 2. The dependence of the q^2 (a) and SDEP (b) performances on the grid mesh. The numbers shown indicates the complexity (an optimal number of components) in the PLS models, respectively.

improved by Motoc.[31] In this method a molecule is separated into partitions of spatial regions either filled or unfilled by atoms or groups of atoms of certain volumes. A similar method was used by Hopfinger to develop 4D-QSAR formalism. An interesting feature of the sector models is their fuzziness.[32] In Figure 1 we show a single-molecule-defined molecular surface that is subsequently replaced for the cubic-like molecular boundary. In such a representation molecular configuration is represented not by entities with sphere-like boundaries but by a polycube formed of a set of cubic domains (cells). Now a shape of the molecule depends on the grid resolution. This also means that two molecules can be compared with the different tolerance, which resembles fuzziness of the SOM neurons of the adjustable tolerance.[32]

Steroids that are complexed by the corticosteroid binding globulin (CBG) are used as a benchmark series in many publications aimed at 3D-QSAR modeling. A number of errors in the steroid structures that can be found in early publications have been corrected in recent years. For a review see ref 33. Similarly to previous studies, we validated the performance of the new method by the leave-one-out cross-validation (LOO CV). Moreover, the steroid series are split into two subseries as previously. In the current study we used the sampling scheme that is usually reported in the literature, i.e., training set: **s1**−**s21**; test set: **s22**−**s31**. Within the grid mesh of 1−1.5 Å the performance of the s-CoMSA only slightly depends on the resolution, as shown in Figure 2, while the LOO CV q^2 takes a value of ca. 0.90−0.88 (for compounds **s1**−**s21**; SDEP = 0.78−0.73 for compounds **s22**−**s31**). This compares advantageously to the CoMFA performances (q^2 = 0.73 for compounds **s1**−**s21**,[33] SDEP − compounds **s22**−**s31** − 0.837[34]). On the other hand, this resembles the performance values obtained in the neural version of the SOM-CoMSA (q^2 = 0.88 for all compounds; SDEP for compounds **s22**−**s31** − 0.69)[8] or this reported for Quasar calculations (q^2 = 0.90 − for compounds **s1**−**s21**).[35]

Although the elimination of variables is always risky, this can both improve the performance of the PLS model and indicate such areas of the molecules that particularly contribute to the activity. Figure 3a,b shows the q^2 (for the training set molecules **s1**−**s21**) and SDEP (for the test set molecules **s22**−**s31**) performances during the IVE-PLS procedure as a function of the number of variables eliminated. The q^2 and SDEP performances depend not only on the number of variables eliminated but also on the complexity, i.e., a number of the PLS components included in the respective models. This effect can be observed especially for the SDEP values, for example, in Figure 2 illustrating the performances for the different models resulted from the different grid resolution. It is evident that a model of the

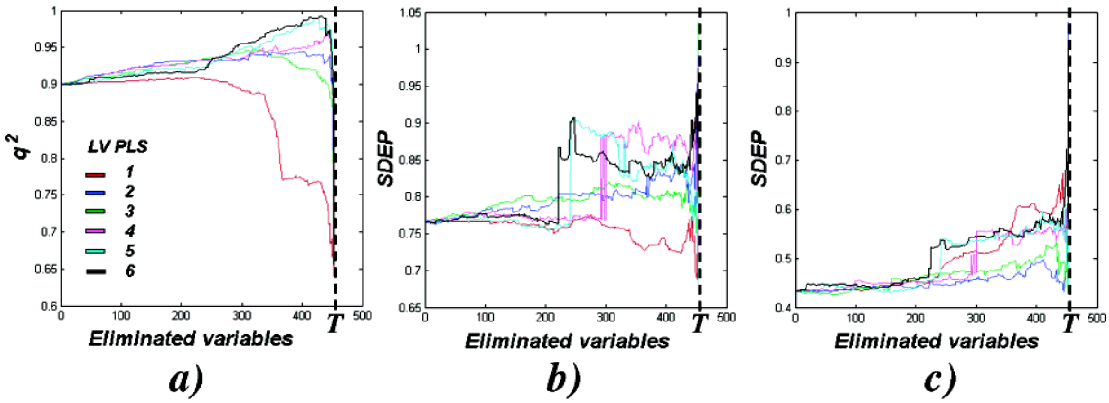


Figure 3. The dependence of the q^2 for the training set **s1**−**s21** (a) and SDEP performances for the test set **s22**−**s31** (b) or **s22**−**s30** (c) during IVE-PLS variable elimination upon the number of variable eliminated. Different colors illustrate the models of the different complexities (details in text), and T indicates a total number of original variables (nonempty grid cells).

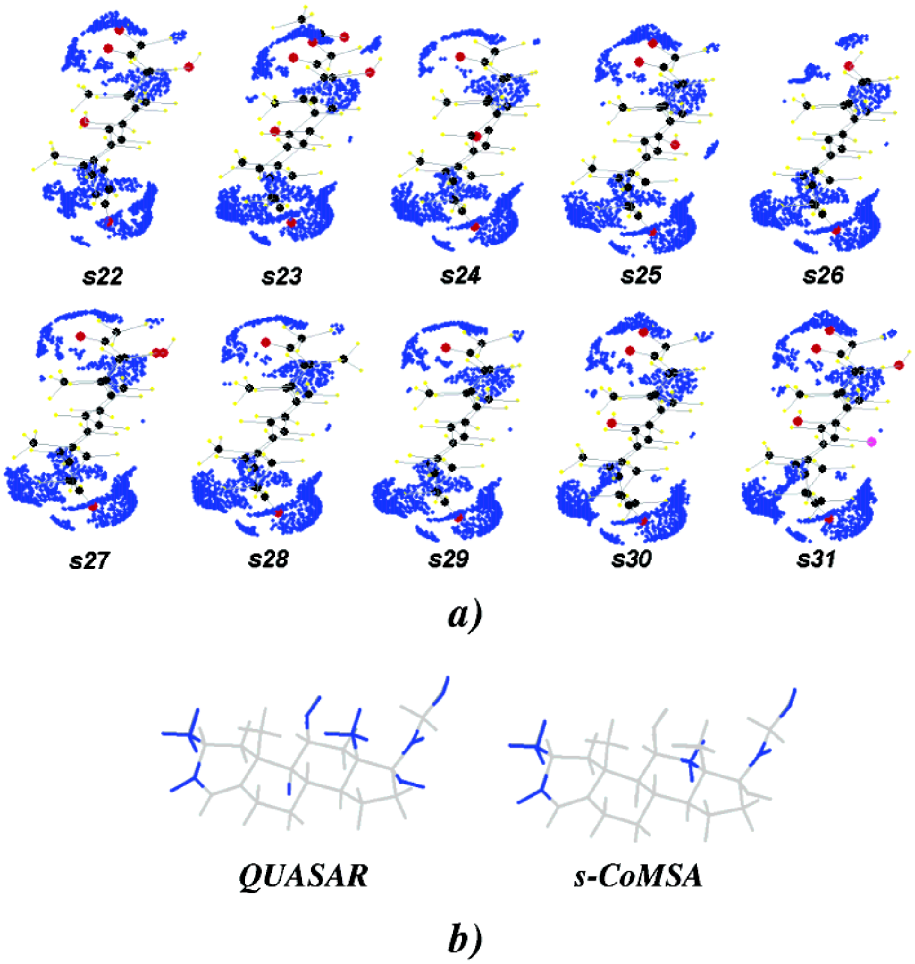


Figure 4. The surface areas (illustrated by the points sampled) of the highest contribution to the CBG activity, as indicated by the IVE-PLS performed for the training set **s1**−**s21**, illustrated for the test set molecules **s22**−**s31** (a), and compared to the respective Quasar results (b).[35]

highest complexity (5) provides also the lowest SDEP predictivity. Therefore, the IVE-PLS procedure was performed in such a way that a number of the PLS latent components were always optimized. This number, however, was truncated not to exceed a value of 1−6, respectively. This means that model complexity cannot exceed 1 in the plot shown in red, while e.g. the plot shown in black (6 components) can include the models of the different

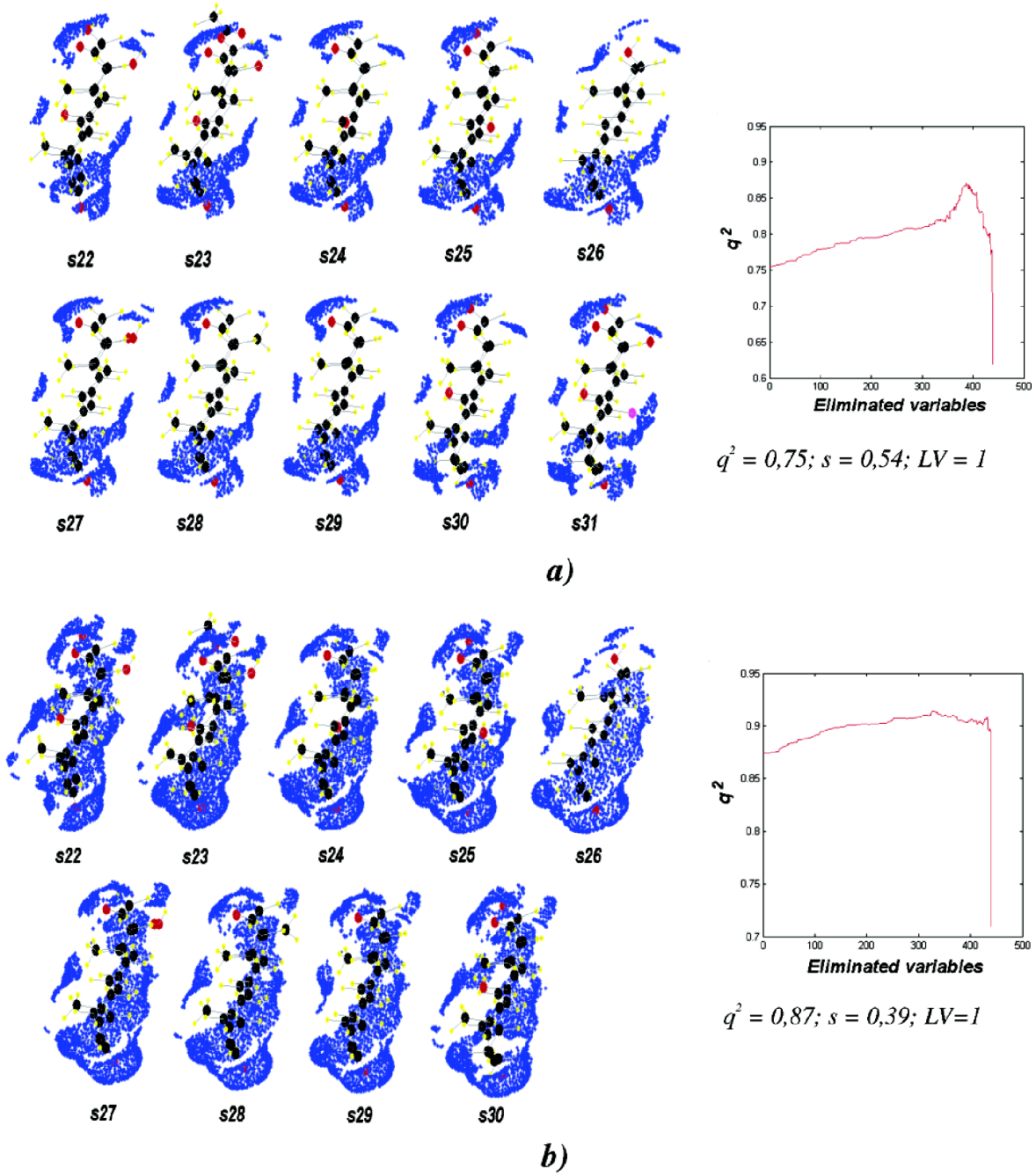


Figure 5. The surface areas (illustrated by the points sampled) of the highest contribution to the CBG activity, as indicated by the IVE-PLS performed for the whole series **s1**−**s31** (a) or molecules **s1**−**s30** (b). The plots shown report the q^2 IVE-PLS performances during variable elimination, respectively. The q^2 and s performances refer to the initial PLS model (before IVE-PLS) and LV indicates a final model complexity. The maximal complexity of the model was truncated to 3.

complexities (each time the optimal value is sought after) from 1 to 6. The complexity of the initial PLS models (elimination was always started from the same model) before variable elimination amounts to 3. Figure 3c illustrates a fact that the elimination of the molecule **s31** from the test set significantly improves the SDEP values. Similarly, to the number of the other methods[33] s-CoMSA evidently misclassifies this compound.

Although for the models of the highest possible complexity (6), variable elimination enables to achieve an impressive value of the q^2 higher than 0.99, the most stable and lowest SDEP values are given for a PLS model with a single component (SDEP = 0.69). This indicates that the models of lower complexity, even if this complexity is lower than the optimal one, are more flexible in handling external objects. However, the evaluation of variable elimination

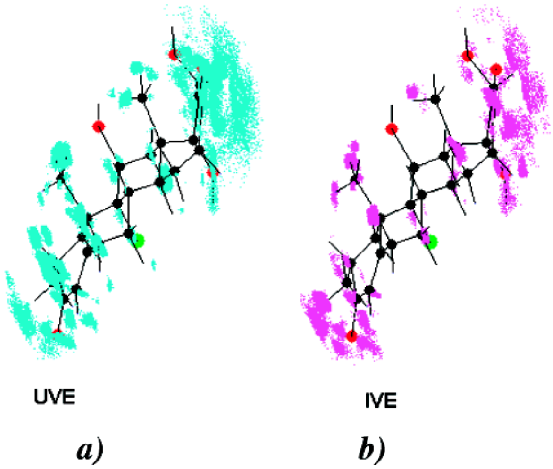


Figure 6. Atomic coordinates of the highest contribution to the CBG activity as indicated by 4D-QSAR−PLS simulations with UVE (a) and IVE (b) data elimination, respectively (details in text). The calculation were conducted using the absolute descriptors occupancy (A_0).

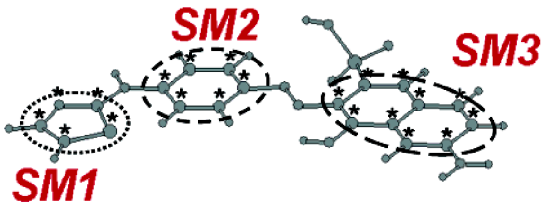


Figure 7. Different superimposition modes SM1−SM3 tested for the dye molecules. The circles indicate the molecule areas covered in the individual superimposition, and the asterisks show individual atoms specified for covering.

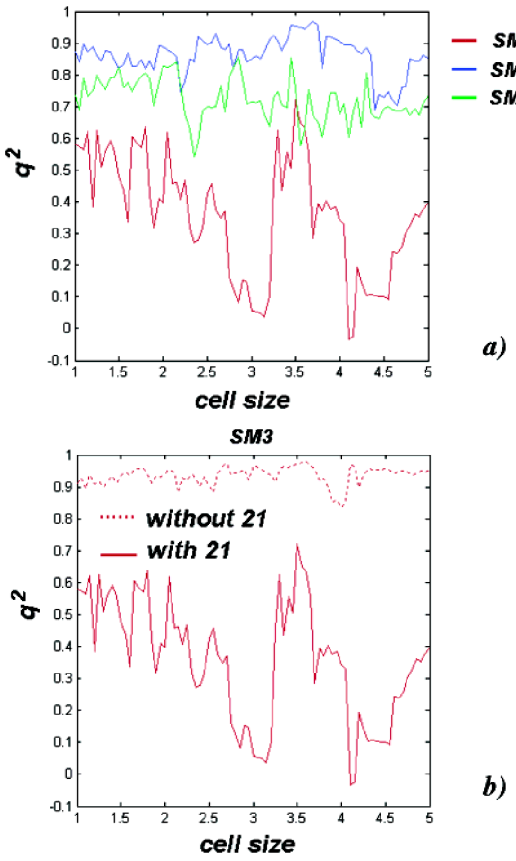


Figure 8. The dependence of the LOO−CV q^2 performance for the s-CoMSA models of the tinctorial properties (**d1**−**d21**) upon the grid resolution for three different superimpostion modes SM1−SM3 tested (a). The exclusion of the molecule **d21** can significantly improve the q^2 performance (b).

without **s31** (Figure 3c) indicates that a 2- or 3-component model gives better predictivity (lower SDEP values). Thus, we used such a model for pointing the possible pharmacophoric sites on the molecular surface of the test set molecules. In Figure 4a we compare the structures of the steroids **s22**−**s31** showing the points sampled originally on the molecular surface and located within the grid volumes that survived after IVE-PLS variable elimination in the training set **s1**−**s21**, as reported in Figure 3. Figure 4b indicates clear similarities between these regions and the ones visualized by the Quasar method.[35]

Alternatively, Figure 5 illustrates similar variable elimination procedure performed, however, for the whole compound series **s1**−**s31** (Figure 5a) or **s1**−**s30** (Figure 5b). The plots shown report the q^2 values during variable elimination. Unlike previously, now the procedure cannot be additionally monitored by the calculation of the SDEP, because all compounds are used during the modeling step. Only compounds **s22**−**s31** or **s22**−**s30** are shown in Figure 5 in order for the easier comparison with Figure 4. The comparison of Figures 4 and 5 illustrates that a change of the compound series used in the IVE-PLS affects the results obtained, i.e., different surface areas are selected, which emphasizes the stochastic nature of this procedure. Moreover, it can be observed that the molecular basis of **s1**−**s30** approves the largest number of variables into a final model. This effect results from a fact that in this case the IVE-PLS starts from a relatively high q^2 level. Moreover, the q^2 increase is much less steep, and a large plateau can be observed near the maximal q^2 value. Thus, in this particular case the procedure is much more stable, and even if we go beyond the maximal q^2 value any sharp decline in the model q^2 cannot be observed. However, the maximal q^2 complies with quite a large number of original variables as shown in Figure 5b.

A question on the relative performance of the s-CoMSA and 4D-QSAR methods may arise. Steroid series seems to be a proper object for the comparison. Therefore, we performed 4D-QSAR simulations with UVE (version modified according to ref 11) and IVE-PLS. The best performance ($q^2 = 0.94$, $s = 0.38$, SDEP = 0.77, including **s31**) compares well to that of the s-CoMSA and are significantly better than this reported for 4D-QSAR without data elimination ($q^2 = 0.84$, $s = 0.50$, SDEP = 0.83, including **s31**).[36] Similarly, the molecular areas (atomic coordinates) indicated as important for the activity in 4D-QSAR, as illustrated in Figure 6, resemble those suggested by s-CoMSA (surfaces defined by respective atoms). This fact seems to show that for the relatively rigid steroid structures the investigations of the conformational space by 4D-QSAR calculation does not improve the performance, and these calculations are extremely time-consuming processes.

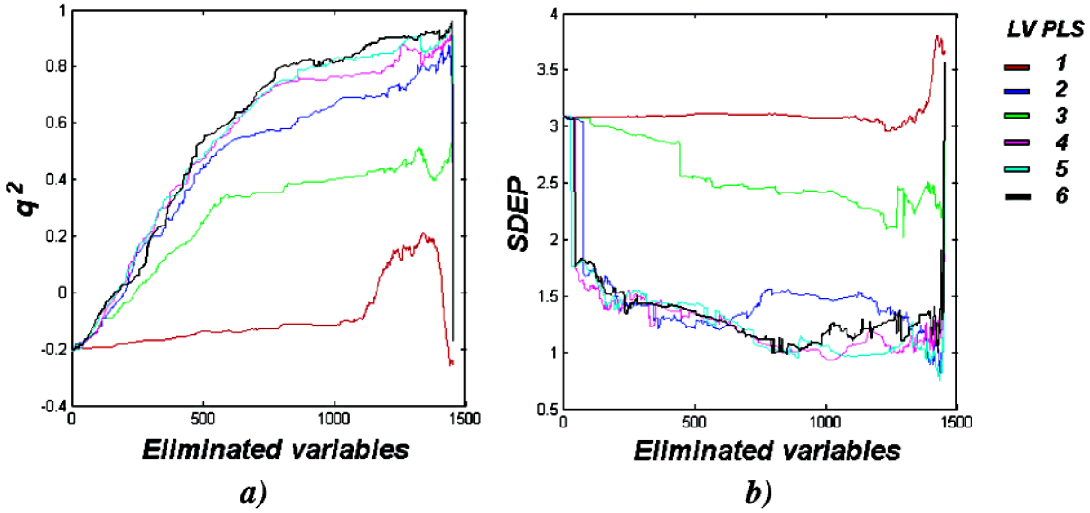


Figure 9. The dependence of the q^2 for the training set **d1**:2:**d21** (a) and SDEP performances for the test set **d2**:2:**d20** (b) during IVE-PLS variable elimination upon the number of variable eliminated. Different colors illustrate the models of the different complexities (details in text).

The interaction between a dye molecule and cellulose is a complicated phenomenon, which can be described by the Langmuir isotherm.[37,38] An isotherm does not, however, provide a molecular description of the process. Moreover, we cannot use such an approach for the optimization of the molecular structure of dye. The influence of the electrostatic, van der Waals, or hydrogen bonding as well as hydrophobic forces on the dyeing process has been investigated.[37] On the other hand, it has been speculated that specific binding sites exist on the crystalline region of the supramolecular cellulose structure that forms holes and cavities capable of incorporating a dye molecule.[39] Does this mean that a similarity between the drug−receptor and dye−fiber interactions makes possible to extend a pharmacophore concept and develop an idea of *tinctophore* in dye chemistry to predict tinctorial properties of dyes by the use of QSAR or related methods? Although it is not clear whether we can treat it similarly to the contacts taking place during targeting a receptor by a drug molecule, several QSAR studies have been published recently[13,40−47] that make use of this concept in investigations of cellulose dyeing. Both 2D- and 3D-QSAR modeling have been applied, including the Hansch, MTD, and Comparative Molecular Field Analysis (CoMFA) methods that appeared to provide quite satisfactory models for different compound series.[23,44−46] In particular, the results of the CoMFA and CoMSA methods indicated that the electrostatic field predominates. On the other hand, dye−cellulose interactions seemed to be less specific than drug−receptor interactions.[12,13,38]

Below the results of the application of s-CoMSA for the analysis of the heterocyclic dye series are reported. Figure 7 indicates three different superposition modes SM1−SM3 tested. In Figure 8 we show the LOO CV q^2 value as a function of the grid resolution. Independent of the grid size, superposition SM3 gives lower q^2 performances than the SM1 and SM2 ones. On the other hand, it can be observed that the exclusion of a single compound **d21** can improve the SM3 q^2 value. This gives similar values to those of the SM1 or SM2 (Figure 8b). In Figure 9 we reported the results of the data reduction experiment that is performed under the control of both the values of q^2 − for the training set (compounds **d1**, **d3**, **d5**, **...**, **d21**) and SDEP − for the test set (compounds **d2**, **d4**, **d6**, **...**, **d20**). The experiment started from a very low initial q^2 value (−0.2). However, data reduction with the IVE-PLS procedure can improve this value up to more then 0.9. Unlike for the benchmark steroid series, now the models of higher complexities provide better models both in respect to the q^2 and SDEP values. Figure 10 indicates the molecular areas indicated by the IVE-PLS variable elimination as important for the tinctorial properties of the dye series. The comparison of these results to those obtained with the SOM-CoMSA version indicates the neural method is better at giving a q^2 value of 0.98.[13] Moreover, the SOM-CoMSA indicates for the SM2 region (best models are obtained for the templates based on central part of the

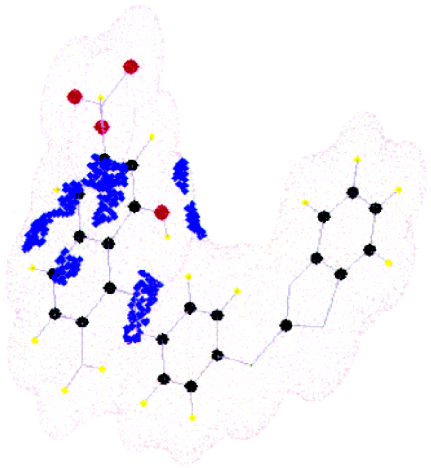


Figure 10. The surface areas (illustrated by the points sampled) of the highest contribution to the tinctorial properties, as indicated by the IVE-PLS procedure.

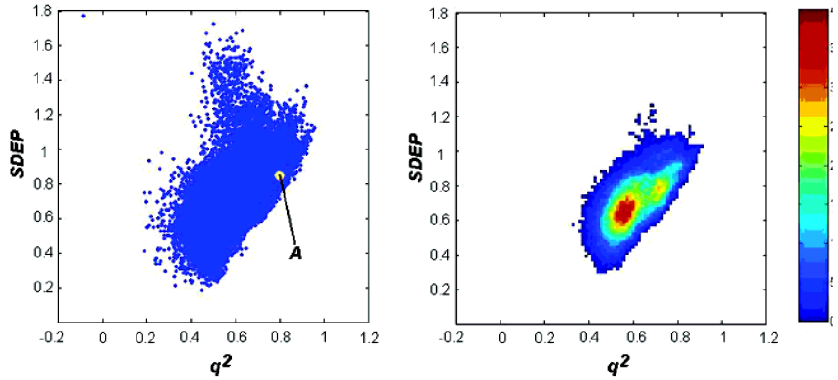


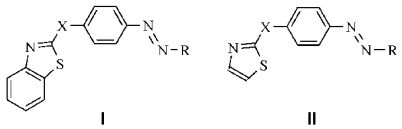
Figure 11. The influence of the sampling of the compounds into the training and test set for the CoMFA calculations of the steroid CoMFA series. The relationships between the LOO−CV q^2 performance estimated for all possible 21-molecule-containing training sets sampled randomly from **s1**−**s31** and SDEP calculated for the 10-molecule-conteining test sets, respectively. The results are shown as a standard plot (a) or a density colormap (b). A indicates the results for the literature sampling—training: **s1**−**s21**; test: **s22**−**s31**.[48]

molecule),[13] i.e., similar areas to those that survive IVE-PLS variable elimination in the s-CoMSA method.

We bear in mind a fact that very low q^2 performance was obtained if we sampled only a half of the dye molecules (**d1**−**d21**) into the training set. We used here a random sampling with the molecules **d1**, **d3**, **d5**, **...**, **d21** (training) and **d2**, **d4**, **d6**, **...**, **d20** (test). Thus an attempt to model 3D-QSAR for such a training set gives a q^2 value of ca. −0.2, while for the whole series a value of 0.9 was obtained. Indeed, we have carefully analyzed the influence of the sampling of the compounds into the training and test series in 3D-QSAR modeling. It was discovered that q^2 significantly depends on such sampling. Although the detailed discussion of this effect is beyond the scope of this paper we would like to indicate here an example of the CoMFA calculation for the steroid series **s1**−**s31**. In the experiment performed we sampled all 31 molecules into the training series including 21 molecules and test series (10 molecules).[48] The total amount of different sampling in such an experiment amounts to 44 352 165 (31!/21!*10!). As it is technically impossible to verify all these models we tested each 1 of 500 possible, which makes 8 870 433 different combinations (Figure 11). In this experiment the q^2 performance ranges from −0.16 to 0.95. Similar SOM-CoMSA calculations provide significantly better performances of $q^2 = 0.42-0.98$. It is worth mentioning that the effect described does not indicate the instability of our 3D-QSAR calculations; we obtained a standard q^2 value of 0.79 and SDEP = 0.83 for the CoMFA model with literature sampling of training/test **s1**−**s21/s22**−**s31** but emphasizes the stochastic nature of the 3D-QSAR calculations performed for a few molecular objects using highly multidimensional molecular descriptor data.

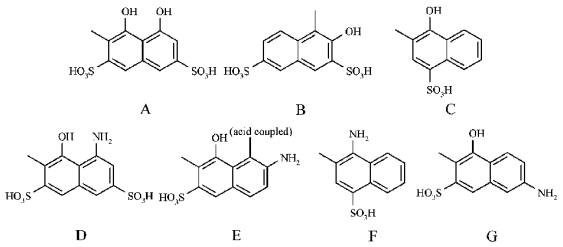
To further test the s-CoMSA method we performed a 3D-QSAR study of a series of HEPT derivatives, inhibitors of the reverse transcriptase enzyme of the HIV virus. 3D-QSAR investigations for a large group of 107 HEPT analogues or different subgroups have been recently reported in many papers.[24,49−55] The reported q^2 performances ranged from $q^2 = 0.92$, $s = 0.36$ for 12 compounds[49] to $q^2 = 0.78$;[24] $q^2 = 0.82$ (in both studies the s value has not been given)[51] or $q^2 = 0.86$.[52] The CoMFA analysis for a series of 101 molecules gives for a training set a performance of $q^2 = 0.86$, $s = 0.53$ (80 compounds, one outlier eliminated).[50] The s-CoMSA

Table 2. Heterocyclic Dye Structures and the Affinity Data to Cellulose[23]



I II

no.		X	R	$-\Delta\mu^0$ (kJ/mol)
d1	I	-NH-	A	6.78
d2	I	-NH-	B	9.20
d3	I	-NH-	D	12.60
d4	I	-NH-	E	15.30
d5	I	O	A	3.26
d6	I	O	B	5.27
d7	I	O	D	7.61
d8	I	O	E	10.30
d9	I	O	G	10.20
d10	I	S	A	1.26
d11	I	S	B	3.56
d12	I	S	D	5.02
d13	I	S	E	8.45
d14	I	S	G	8.12
d15	II	-NH-	E	15.33
d16	II	-NH-	D	12.60
d17	II	-NH-	B	9.24
d18	II	-NH-	A	6.80
d19	I	S	C	5.86
d20	I	S	F	10.33
d21	I	S	E (acid coupled)	9.75

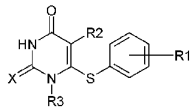


A B C

D E F G

model obtained in our study significantly outperformed these values. Thus, the performance of the optimal s-CoMSA IVE-PLS LOO−CV model obtained for the grid mesh of 1.5 Å amounts to $q^2 = 0.86$, s = 0.58 for the whole series of 107 compounds **h1**−**h107**. To validate the model predictivity we divided the series into the training (**h1**−**h80**) and test set (**h81**−**h107**), i.e., using similar rules as those reported

Table 3. Chemical Structures with the Observed[24] and Calculated Values of Anti-HIV Activity for the HEPT Derivatives



no.	R1	R2	R3	X	obs	pred[a]	pred[b]	pred[c]	pred[c]	pred[d]
h1	2-Me	Me	$CH_2OCH_2CH_2OH$	O	4.15	4.75	4.47[e]	4.47	5.15	3.84
h2	2-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.85	5.03	3.74	3.76	3.83	4.10
h3	2-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.72	4.11	3.40	4.65	5.03	4.94
h4	3-Me	Me	$CH_2OCH_2CH_2OH$	O	5.59	5.40	4.70[e]	4.83	4.97	5.44
h5	3-Et	Me	$CH_2OCH_2CH_2OH$	O	5.57	5.33	4.73[e]	5.11	4.94	5.65
h6	3-t-Bu	Me	$CH_2OCH_2CH_2OH$	O	4.92	5.30	5.03[e]	5.36	4.97	4.93
h7	3-CF_3	Me	$CH_2OCH_2CH_2OH$	O	4.35	5.21	4.71	4.09	4.39	4.62
h8	3-F	Me	$CH_2OCH_2CH_2OH$	O	5.48	4.67	5.11	4.54	4.80	5.29
h9	3-Cl	Me	$CH_2OCH_2CH_2OH$	O	4.89	5.08	5.00[e]	4.86	4.98	5.26
h10	3-Br	Me	$CH_2OCH_2CH_2OH$	O	5.24	5.09	4.81[e]	5.08	4.99	5.24
h11	3-I	Me	$CH_2OCH_2CH_2OH$	O	5.00	5.29	4.80	5.64	5.23	5.26
h12	3-NO_2	Me	$CH_2OCH_2CH_2OH$	O	4.47	4.79	5.37	4.62	4.33	4.57
h13	3-OH	Me	$CH_2OCH_2CH_2OH$	O	4.09	5.34	5.28	4.72	4.71	4.93
h14	3-OMe	Me	$CH_2OCH_2CH_2OH$	O	4.66	5.02	5.55	5.50	5.07	5.23
h15	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	O	6.59	5.99	5.90[e]	6.27	6.44	6.42
h16	3,5-Cl_2	Me	$CH_2OCH_2CH_2OH$	O	5.89	5.88	6.07[e]	6.25	6.21	6.28
h17	3,5-Me_2	Me	$CH_2OCH_2CH_2OH$	S	6.66	6.37	6.21	6.60	6.33	6.50
h18	3-COOMe	Me	$CH_2OCH_2CH_2OH$	O	5.10	4.83	3.77	5.37	4.83	4.63
h19	3-COMe	Me	$CH_2OCH_2CH_2OH$	O	5.14	4.70	4.69	5.55	5.12	4.36
h20	3-CN	Me	$CH_2OCH_2CH_2OH$	O	5.00	4.77	5.59	5.27	5.08	4.72
h21	H	$CH_2CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	5.60	5.61	5.47	5.60	5.18	5.68
h22	H	Et	$CH_2OCH_2CH_2OH$	S	6.96	6.24	6.27	6.35	6.92	6.74
h23	H	Pr	$CH_2OCH_2CH_2OH$	S	5.00	6.71	6.4[e]	6.79	5.88	6.01
h24	H	i-Pr	$CH_2OCH_2CH_2OH$	S	7.23	7.58	7.88	6.75	6.15	7.32
h25	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	S	8.11	7.32	6.97[e]	7.48	7.69	7.76
h26	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	S	8.30	8.51	8.28	8.45	8.26	8.30
h2	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	S	7.37	7.26	6.54	8.12	7.84	7.64
h28	H	Et	$CH_2OCH_2CH_2OH$	O	6.92	6.21	5.50[e]	5.97	6.85	6.66
h29	H	Pr	$CH_2OCH_2CH_2OH$	O	5.47	6.43	5.52[e]	6.07	5.43	5.93
h30	H	i-Pr	$CH_2OCH_2CH_2OH$	O	7.20	7.11	6.27	6.79	6.83	7.24
h31	3,5-Me_2	Et	$CH_2OCH_2CH_2OH$	O	7.89	7.52	6.41	7.24	7.79	7.68
h32	3,5-Me_2	i-Pr	$CH_2OCH_2CH_2OH$	O	8.57	8.30	7.45	8.20	8.55	8.22
h33	3,5-Cl_2	Et	$CH_2OCH_2CH_2OH$	O	7.85	7.25	6.32	7.51	7.84	7.56
h34	4-Me	Me	$CH_2OCH_2CH_2OH$	O	3.66	4.90	4.52[e]	5.36	3.71	5.39
h35	H	Me	$CH_2OCH_2CH_2OH$	O	5.15	5.10	4.62[e]	4.90	5.04	5.30
h36	H	Me	$CH_2OCH_2CH_2OH$	S	6.01	5.42	5.92	5.26	5.27	5.38
h37	H	I	$CH_2OCH_2CH_2OH$	O	5.44	5.64	5.11[e]	5.68	5.41	5.66
h38	H	$CH{=}CH_2$	$CH_2OCH_2CH_2OH$	O	5.69	6.20	5.84	4.80	5.32	6.53
h39	H	CH=CHPh	$CH_2OCH_2CH_2OH$	O	5.22	5.35	6.18	5.40	5.22	4.75
h40	H	CH_2Ph	$CH_2OCH_2CH_2OH$	O	4.37	5.43	5.68	5.28	5.08	4.81
h41	H	$CH{=}CPh_2$	$CH_2OCH_2CH_2OH$	O	6.07	5.11	4.22	4.92	6.10	6.06
h42	H	Me	$CH_2OCH_2CH_2OMe$	O	5.06	4.64	4.71	5.26	5.14	5.35
h43	H	Me	$CH_2OCH_2CH_2OAc$	O	5.17	4.98	4.71	5.24	5.05	4.62
h44	H	Me	$CH_2OCH_2CH_2OCOPh$	O	5.12	4.59	4.78	5.53	5.24	5.66
h45	H	Me	CH_2OCH_2Me	O	6.48	5.76	5.88	5.82	5.75	5.75
h46	H	Me	$CH_2OCH_2CH_2Cl$	O	5.82	5.99	6.02	5.87	5.84	5.73
h47	H	Me	$CH_2OCH_2CH_2N_3$	O	5.24	5.71	4.99	6.07	5.18	4.74
h48	H	Me	$CH_2OCH_2CH_2F$	O	5.96	4.94	6.27	5.49	5.39	5.23
h49	H	Me	$CH_2OCH_2CH_2Me$	O	5.48	5.72	6.08[e]	5.60	5.26	5.67
h50	H	Me	CH_2OCH_2Ph	O	7.06	6.25	6.92	6.88	6.96	6.42
h51	H	Et	CH_2OCH_2Me	O	7.72	6.96	6.73[e]	6.72	6.80	7.14
h52	H	Et	CH_2OCH_2Me	S	7.58	6.92	7.16	6.91	6.79	7.22
h53	3,5-Me_2	Et	CH_2OCH_2Me	O	8.24	8.35	8.03	8.12	8.24	8.17
h54	3,5-Me_2	Et	CH_2OCH_2Me	S	8.30	8.27	7.64	8.29	8.21	8.26
h55	H	Et	CH_2OCH_2Ph	O	8.23	7.78	7.73[e]	6.63	7.41	7.64
h56	3,5-Me_2	Et	CH_2OCH_2Ph	O	8.55	9.74	9.21	8.54	8.36	8.51
h57	H	Et	CH_2OCH_2Ph	S	8.09	7.06	7.38[e]	7.53	8.09	7.72
h58	3,5-Me_2	Et	CH_2OCH_2Ph	S	8.14	9.00	7.66	8.75	8.17	8.59
h59	H	i-Pr	CH_2OCH_2Me	O	7.99	7.48	7.48	7.67	8.13	7.72
h60	H	i-Pr	CH_2OCH_2Ph	O	8.51	8.44	8.49	8.53	8.19	8.16
h61	H	i-Pr	CH_2OCH_2Me	S	7.89	8.07	8.32	7.84	7.79	7.80
h62	H	i-Pr	CH_2OCH_2Ph	S	8.14	7.78	8.52	8.31	8.12	8.24
h63	H	Me	CH_2OMe	O	5.68	5.85	5.87[e]	5.90	5.68	6.08
h64	H	Me	CH_2OBu	O	5.33	5.89	6.07[e]	5.73	5.64	5.58
h65	H	Me	Et	O	5.66	6.17	5.06	6.29	5.51	6.64
h66	H	Me	Bu	O	5.92	5.42	4.84	6.41	5.61	5.98
h67	3,5-Cl_2	Et	CH_2OCH_2Me	S	7.89	8.30	7.18	8.00	7.81	8.13
h68	H	Et	$CH_2O{-}$i-Pr	S	6.66	7.06	7.19	6.79	6.64	6.87

Table 3 (Continued)

no.	R1	R2	R3	X	obs	pred[a]	pred[b]	pred[c]	pred[c]	pred[d]
h69	H	Et	CH_2O−c-Hex	S	5.79	5.85	7.00	6.63	5.95	6.06
h70	H	Et	CH_2OCH_2-c-Hex	S	6.45	7.13	7.53	6.57	7.06	6.01
h71	H	Et	$CH_2OCH_2C_6H_4$(4-Me)	S	7.11	7.56	7.45	6.97	7.63	7.57
h72	H	Et	$CH_2OCH_2C_6H_4$(4-Cl)	S	7.92	7.01	7.33	7.44	7.99	7.63
h73	H	Et	$CH_2OCH_2CH_2Ph$	S	7.04	6.75	7.50	6.61	6.82	7.60
h74	3,5-Cl_2	Et	CH_2OCH_2Me	O	8.13	8.41	8.22	8.03	8.21	8.05
h75	H	Et	CH_2O−i-Pr	O	6.47	6.46	6.01	6.79	6.71	6.78
h76	H	Et	CH_2O−c-Hex	O	5.40	6.09	6.46	5.80	5.18	5.98
h77	H	Et	CH_2OCH_2-c-Hex	O	6.35	7.32	6.93[e]	6.44	6.45	5.93
h78	H	Et	$CH_2OCH_2CH_2Ph$	O	7.02	7.66	6.88	7.10	7.16	7.52
h79	H	c-Pr	CH_2OCH_2Me	S	7.02	7.32	6.61	7.39	6.85	6.61
h80	H	c-Pr	CH_2OCH_2Me	O	7.00	6.95	6.28[e]	7.16	7.01	6.53
h81	H	Me	$CH_2OCH_2CH_2OC_5H_{11}$-n	O	4.46	5.05	4.84[e]	6.01	5.24	5.12
h82	2-Cl	Me	$CH_2OCH_2CH_2OH$	O	3.89	4.37	4.27	4.64	5.14	4.31
h83	3-CH_2OH	Me	$CH_2OCH_2CH_2OH$	O	3.53	6.62	5.26	4.39	4.83	4.60
h84	4-F	Me	$CH_2OCH_2CH_2OH$	O	3.60	5.00	3.80	4.50	3.58	5.10
h85	4-Cl	Me	$CH_2OCH_2CH_2OH$	O	3.60	5.33	4.00[e]	5.04	3.80	5.45
h86	4-NO_2	Me	$CH_2OCH_2CH_2OH$	O	3.72	6.29	3.91	4.43	4.22	5.04
h87	4-CN	Me	$CH_2OCH_2CH_2OH$	O	3.60	5.80	3.18	5.65	4.17	4.98
h88	4-OH	Me	$CH_2OCH_2CH_2OH$	O	3.56	5.08	4.20	4.95	3.57	4.74
h89	4-OMe	Me	$CH_2OCH_2CH_2OH$	O	3.60	5.83	4.15	4.98	3.61	5.28
h90	4-COMe	Me	$CH_2OCH_2CH_2OH$	O	3.96	7.04	3.54	4.91	3.68	5.00
h91	4-COOH	Me	$CH_2OCH_2CH_2OH$	O	3.45	6.19	4.31	4.69	4.56	4.67
h92	3-$CONH_2$	Me	$CH_2OCH_2CH_2OH$	O	3.51	5.09	5.33	4.99	4.68	4.14
h93	H	COOMe	$CH_2OCH_2CH_2OH$	O	5.18	8.25	4.76	4.22	5.01	5.28
h94	H	CONHPh	$CH_2OCH_2CH_2OH$	O	4.74	7.15	5.26	5.29	5.13	4.60
h95	H	SPh	$CH_2OCH_2CH_2OH$	O	4.68	7.27	6.10	5.01	5.57	4.97
h96	H	CCH	$CH_2OCH_2CH_2OH$	O	4.74	6.90	4.70[e]	5.95	5.23	6.30
h97	H	CCPh	$CH_2OCH_2CH_2OH$	O	5.47	6.67	4.38	5.86	5.26	4.71
h98	3-NH_2	Me	$CH_2OCH_2CH_2OH$	O	3.60	4.89	4.81	4.76	5.02	4.57
h99	H	$COCHMe_2$	$CH_2OCH_2CH_2OH$	O	4.92	7.80	5.78	5.78	4.58	5.27
h100	H	COPh	$CH_2OCH_2CH_2OH$	O	4.89	6.44	4.00	5.08	5.15	5.11
h101	H	CCMe	$CH_2OCH_2CH_2OH$	O	4.72	6.80	4.68	6.09	5.28	5.57
h102	H	F	$CH_2OCH_2CH_2OH$	O	4.00	6.04	4.61	5.25	5.11	5.00
h103	H	Cl	$CH_2OCH_2CH_2OH$	O	4.52	5.84	4.51[e]	5.46	5.16	5.36
h104	H	Br	$CH_2OCH_2CH_2OH$	O	4.70	5.40	4.66	5.49	5.16	5.47
h105	H	Me	$CH_2OCH_2CH_2OCH_2Ph$	O	4.70	4.61	4.68	5.61	5.63	5.93
h106	H	Me	H	O	3.60	5.80	6.18	6.06	5.52	5.63
h107	H	Me	Me	O	3.82	4.64	4.90	6.16	5.49	.75

[a] s-CoMSA. [b] s-CoMSA for the Kennard−Stone training/test set sampling. [c] Activity values according to the ref 24. [d] Activity values according to the ref 51. [e] Test set.

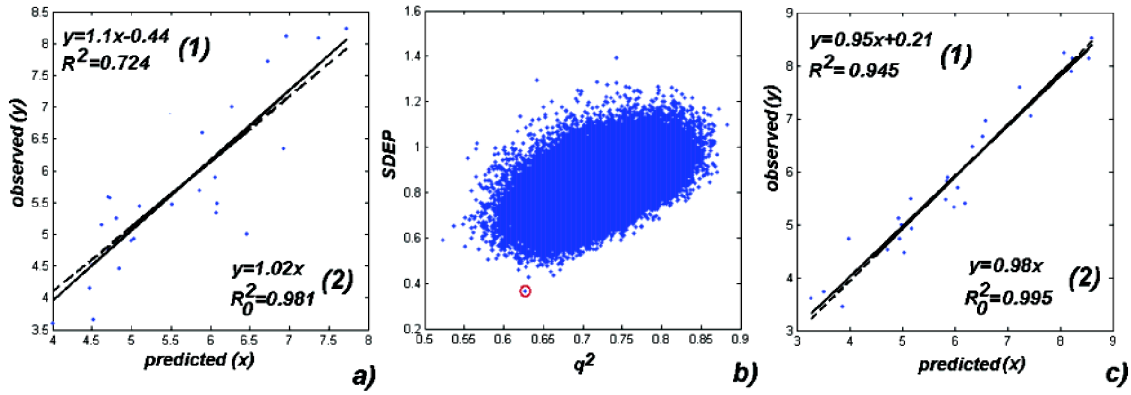


Figure 12. The GT model validation for the Kennard−Stone training/test set sampling (a), the dependence of the SDEP vs q^2 for 110 000 random training/test set samplings (b), and the GT model validation for the highly predictive model selected among these models (SDEP = 0.37) (c). We tried to keep the style of the presentation of the authors.[56] The regression between observed (Y) and predicted (X) activity values for the test set. The solid line shows the regressional equation given by (1). The dotted line illustrates the regression without the bias (2). The closer are these linear plots, the better is the model predictivity. Calculations after $\text{pred}_i^0 = k^*\text{pred}_i$, $k = \sum\text{obs}_i\text{pred}_i/\sum\text{pred}_i^2$, and $R_0^2 = 1 - \sum(\text{pred}_i - \text{pred}_i^0)^2/\sum(\text{pred}_i - \text{mean(pred)})^2$ where the upper index 0 relates to regression observed (Y) vs predicted (X), k is a slope of the regression through the origin (2), and R_0^2 is the correlation coefficient for the regression of observed (Y) vs predicted (X) without bias. $[(R^2 - R_0^2)/R^2] < 0.1$ and $0.85 \le k \le 1.15$ as recommended by Golbraikh and Tropsha.[56]

previously.[24,51] The predicted values are given in Table 3, respectively. This procedure results in the relatively high SDEP value of 2.01. Moreover, we validated this model performing the calculation of the Golbraikh−Tropsha (GT)

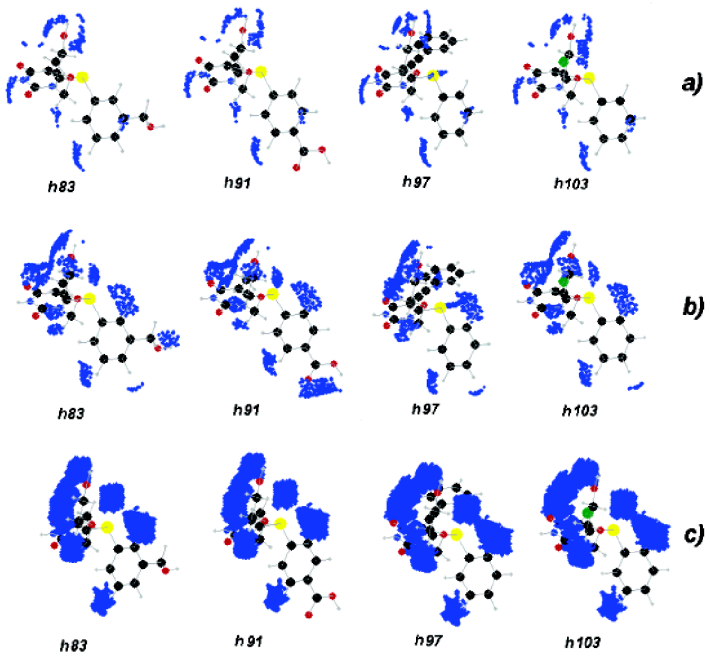


Figure 13. The surface areas (illustrated by the points sampled) of the highest contribution to the anti-HIV activity, as indicated by the IVE-PLS procedure for the training series **h1**−**h80** (a) and the whole series **h1**−**h107** (b). Alternatively, the areas surviving IVE-PLS for the whole series **h1**−**h107** were aggregated for all molecules, and these (25%) of the lowest population were eliminated (c). The plots illustrate only selected molecules.

criterion,[56] as shown in Supporting Information. Clearly, our model does not satisfy this criterion. However, similar results were obtained for the models given in the literature. We think the reason for such behavior is a careless training and test subset selection. Thus, we used the Kennard−Stone (KS) algorithm[57] for sampling the compounds to the training/test subsets (test set: **h1**, **h4**, **h5**, **h6**, **h9**, **h10**, **h15**, **h16**, **h23**, **h25**, **h28**, **h29**, **h34**, **h35**, **h37**, **h49**, **h51**, **h55**, **h57**, **h63**, **h64**, **h77**, **h80**, **h81**, **h85**, **h96**, **h103**). In Figure 12a we illustrated the results of model validation by the GT calculations. Due to the lack of the independent variable (*X*) data used in investigations reported in the literature we could not repeat this procedure to compare our performance to those represented by the literature data. Model quality is much better now, and SDEP value amounts to 0.69. By random testing (Figure 12b) of the relatively small fraction (110 000 models) of all possible 107!/(80!*27!) samplings of the 107 HEPT molecules into the training and test sets (test set: **h3**, **h6**, **h8**, **h11**, **h16**, **h17**, **h22**, **h29**, **h38**, **h44**, **h46**, **h52**, **h53**, **h58**, **h60**, **h61**, **h62**, **h64**, **h73**, **h75**, **h76**, **h81**, **h86**, **h87**, **h91**, **h101**, **h103**; 80/27 molecules) we were able to find models even better fulfilling the GT criterion (Figure 12c). This clearly indicates that the s-CoMSA method provides reliable and highly predictive models. Figure 13 illustrates the molecular plots indicating the regions important for the activity of some representative molecules.

CONCLUSIONS

Shape analysis is a powerful tool in chemistry and drug design. It seems that the approaches including explicit shape representations should gain more attention, being more adequate for the analysis of molecular phenomena that are usually more of less spatially oriented processes. Thus, for example, the investigations of the molecular recognition or receptor−ligand interactions should probably focus *somewhere near a molecular surface* that is still more precise information than *anywhere near a molecule*. In the current publication we have presented a novel formalism for the comparative molecular surface analysis (s-CoMSA). The method enables both quantitative 3D-QSAR modeling and finding possible pharmacophoric sites. The method provides very predictive models for the CBG activity of the benchmark steroid series, tinctorial properties of the heterocyclic azo dyes and anti-HIV activity of the HEPT series.

ACKNOWLEDGMENT

The authors thank Professor Johann Gasteiger of the University of Erlangen-Nürnberg, BRD both for his valuable discussion and for facilitating access to the programs of CORINA, PETRA, SURFACE, and KMAP. The financial support of the KBN Warsaw, grants no. T08E02820, PBZ KBN - 4T09A 088 25, and 4 T09A 034 24, is gratefully acknowledged.

Supporting Information Available: Validation of different models by the Golbraikh−Tropsha criteria for HEPT analogues. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Cramer, III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(2) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D QSAR models using the 4D QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.
(3) Hahn, M.; Rogers, D. Receptor surface models. *Perspect. Drug Discov. Des.* **1998**, *12/13/14*, 117−133.

(4) Hahn, M. Receptor surface models: 1. Definition and construction. *J. Med. Chem.* **1995**, *38*, 2080−2090.
(5) Hahn, M.; Rogers, D. Receptor surface models: 2. Application to QSAR. *J. Med. Chem.* **1995**, *38*, 2091−2102.
(6) Ivanciuc, O.; Ivanciuc, T.; Cabrol-Bass, D. 3D QSAR with CoRSA. Comparative receptor surface analysis. Application to calcium channel agonists. *Analusis* **2000**, *28*, 637−642.
(7) Jain, A. N.; Koile, K.; Bauer, B.; Chapman, D. Compass: Predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315−2327.
(8) Polanski, J.; Walczak, B. The comparative molecular surface analysis (CoMSA): a novel tool for molecular design. *Comput. Chem.* **2000**, *24*, 615−625.
(9) Anzali, S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J.; Teckentrup, A.; Wagener M. The use of self-organizing neural networks in drug design. *Perspect. Drug Discov. Des.* **1998**, *9/10/11*, 273−299.
(10) Polanski, J.; Gieleciak, R.; Bak, A. The comparative molecular surface analysis (CoMSA) − a nongrid 3D QSAR method by a coupled neural network and PLS system: Predicting pK_a values of benzoic and alkanoic acids. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 184−191.
(11) Polanski, J.; Gieleciak, R. The comparative molecular surface analysis (CoMSA) with modified uninformative variable elimination-PLS (UVE-PLS) method: application to the steroids binding the aromatase enzym. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 656−666.
(12) Polanski, J.; Gieleciak, R.; Wyszomirski, M. Comparative molecular surface analysis (CoMSA) for modeling dye-fiber affinities of the azo and antraquinone dyes. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1754−1762.
(13) Polanski, J.; Gieleciak, R.; Wyszomirski, M. Mapping dye pharmacophores by the comparative molecular surface analysis (CoMSA): application to heterocyclic monoazo dyes. *Dyes Pigm.* **2004**, *62*, 63−78.
(14) Polanski, J.; Gasteiger, J.; Jarzembek, K. Self-Organizing neural networks for screening and development of novel artificial sweetener candidates. *Comb. Chem. High Throughput Screening* **2000**, *3*, 481−495.
(15) Polanski, J.; Gieleciak, R. Comparative molecular surface analysis: a novel tool for drug design and molecular diversity studies. *Mol. Diversity* **2003**, *7*, 45−59.
(16) Polanski, J. Self-organizing neural networks for pharmacofore mapping. *Adv. Drug Deliv. Rev.* **2003**, *55*, 1149−1162.
(17) Kohonen, T. *Self-Organization and Associative Memory*, 3rd ed.; Springer: Berlin, 1989.
(18) Wold. S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(19) Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatsu, K. New molecular surface-based 3D-QSAR method using Kohonen neural network and 3-Way PLS. *Comput. Chem.* **2002**, *26*, 583−589.
(20) Hasegawa, K.; Morikami, K.; Shiratori, Y.; Ohtsuka, T.; Aoki, Y.; Shimma, N. 3D-QSAR study of antifungal N-myristoyltransferase inhibitors by comparative molecular surface analysis. *Chemom. Intell. Lab. Syst.* **2003**, *69*, 51−59.
(21) Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatsu, K. Multi-way PLS modeling of structure−activity data by incorporating electrostatic and lipophilic potentials on molecular surface. *Comput. Biol. Chem.* **2003**, *27*, 381−386.
(22) Zupan, J.; Gasteiger, J. *Neural Networks and drug design for Chemists,* 2nd ed.; VCH: Weinheim, 1999.
(23) Timofei, S.; Fabian, W. M. F. Comparative molecular field analysis (CoMFA) of heterocyclic monoazo dye-fiber affinities. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1218−1222.
(24) Jalali-Heravi, M.; Parastar, F. Use of artificial neural networks in a QSAR study of anti-HIV activity for a large group of HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 147−154.
(25) Match3D program package, available from Professor J. Gasteiger, Computer-Chemie-Centrum, University Erlangen-Nurnberg, Germany. See: http://www2.ccc.uni-erlangen.de.
(26) Sybyl 6.5. program, available from the Tripos Inc., St. Louis, MO, U.S.A. http://www.tripos.com.
(27) HyperChem 5.0 program, available from the HyperCube Inc., Gainesville, FL, U.S.A. http://www.hyper.com.
(28) MATLAB 6.5. program, available from The Mathworks Inc., Natick, MA. http://www.mathworks.com.
(29) Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandeginste, B. M. V.; Sterna, C. Elimination of uninformative variables for multivariate calibration. *Anal. Chim. Acta* **1996**, *330*, 1−17.
(30) Testa, B.; Purcell, W. P. A QSAR study of sulfonamide binding to carbonic anhydrase as test of steric models. *Eur. J. Med. Chem.* **1978**, *13*, 509−514.
(31) Motoc, I. *Molecular Shape Descriptors, in Steric Effects in Drug Design*; Charton, M., Motoc, I., Eds.; Akademie: Berlin, 1983; pp 93−105.
(32) Polanski, J. Molecular shape analysis. In *Handbook of chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Verlag: Weinheim, 2003; pp 302−319.
(33) Coats, E. The CoMFA steroids as a benchmark dataset for development of 3D QSAR methods. *Perspect. Drug Discov. Des.* **1998**, *12/13/14*, 199−213.
(34) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-organizing molecular field analysis: A tool for structure−activity studies. *J. Med. Chem.* **1999**, *42*, 573−583.
(35) User and Reference Manual Quasar 4.0. http://www.biograf.ch.
(36) Polanski, J.; Bak, A. Modeling steric and electronic effects in 3D and 4D-QSAR schemes: Predicting benzoic pK_a values and steroid CBG binding affinities. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2081−2092.
(37) Peters, R. H. *Textile chemistry. The physical chemistry of dyeing*; Elsevier: Amsterdam, 1975; Vol. III.
(38) Timofei, S.; Schmidt, W.; Kurunczi, L.; Simon, Z. A Review of QSAR for dye affinity for cellulose fibres. *Dyes Pigm.* **2000**, *47*, 5−16.
(39) French, A. D.; Battista, O. A.; Cuculo, J. A.; Gray, D. G. In *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th ed.; Wiley: New York, 1993; Vol. 5, p 476.
(40) Timofei, S.: Schmidt, W.; Kurunczi, L.; Simmon, Z.; Sallo, A. A QSAR study of the adsorption by cellulose fibre of antraquinone vat dyes. *Dyes Pigm.* **1994**, *24*, 267−279.
(41) Timofei, S.; Kurunczi, L.; Schmidt, W.; Fabian, W. M. F.; Simon, Z. Structure-affinity binding relationships by principal component regresion analysis of antraquinone dyes. *Quant. Struct. Act. Relat.* **1995**, *14*, 444−449.
(42) Timofei, S.; Kurunczi. L.; Schmidt, W.; Simon, Z. Structure-affinity binding relationships of some 4-aminobenzene derivatives for cellulose fibre. *Dyes Pigm.* **1995**, *29*, 251−258.
(43) Timofei, S.; Kurunczi. L.; Schmidt, W.; Simon, Z. Lipophilicity in dye-cellulose fibre binding. *Dyes Pigm.* **1996**, *32*, 25−42.
(44) Fabian, W. M. F. Timofei, S.; Kurunczi. L. Comparative molecular field analysis (CoMFA), semiempirical (AM1) molecular orbital and multiconformational minimal steric difference (MTD) calculation of antraquinone dye-fibre affinities. *J. Mol. Struct. THEOCHEM* **1995**, *340*, 73−81.
(45) Fabian, W. M. F.; Timofei, S. Comparative molecular field analysis (CoMFA) of dye-fibre affinities II: symmetrical bisazo dyes. *J. Mol. Struct. THEOCHEM* **1996**, *362*, 155−162.
(46) Oprea, T. I.; Kurunczi, L.; Timofei, S. QSAR studies of disperse azo dyes towards the negation of the pharmacophore theory of dye − fibre interaction? *Dyes Pigm.* **1997**, *33*, 41−64.
(47) Funar-Timofei, S.; Schüürmann, G. Comparative molecular field analysis (CoMFA) of anionic azo dye-fiber affinities i: gas-phase molecular orbital descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 788−795.
(48) Polanski, J.; Gieleciak, R.; Bak, A. Probability issues in molecular design: Predictive and modeling ability in 3D-QSAR schemes. *Comb. Chem. High Throughput Screening* in print.
(49) Kireev, D. B.; Chretien, J. R.; Grierson, D. S.; Monneret, C. A 3D QSAR study of a series of HEPT analogues: the influence of conformational mobility on HIV-1 reverse transctriptase inhibition. *J. Med. Chem.* **1997**, *40*, 4257−4264.
(50) Hannongbua, S.; Nivesanond, K.; Lawtrakul, L.; Pungpo, P.; Wolschann, P. 3D-Quantitative structure−activity relationships of HEPT derivatives as HIV-1 reverse trascriptase inhibitors, based on ab initio calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 848−855.
(51) Luco, J. M.; Ferretii, F. H. QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 392−401.
(52) Douali, L.; Villemin, D.; Charquaoui, D. Comparative QSAR based on neural networks for the anti-HIV activity of HEPT derivatives. *Curr. Pharm. Des.* **2003**, *9*, 1817−1826.
(53) Mager, P. P. Hybrid canonical-correlation neural-network approach applied to nonnucleoside HIV-1 reverse transcriptase inhibitors (HEPT derivatives). *Curr. Med. Chem.* **2003**, *10*, 1643−1659.
(54) Douali, L.; Villemin, D.; Charquaoui, D. Neural networks: accurate nonlinear QSAR model for HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1200−1207.
(55) Gayen, S.; Debnath, B.; Samanta, S.; Jha, T. QSAR study on some anti-HIV HEPT analogues using physicochemical and topological parameters. *Bioorg. Med. Chem.* **2004**, *12*, 1493−1503.
(56) Golbraikh. A.; Thropsha, A. Beware of q^2 ! *J. Mol. Graph. Mod.* **2002**, *20*, 269−276.
(57) Kennard, R. W.; Stone, L. A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137−148.

CI049960L

Molecules **2004**, *9*, 1148–1159

molecules
ISSN 1420-3049
http://www.mdpi.org

Self-organizing Neural Networks for Modeling Robust 3D and 4D QSAR: Application to Dihydrofolate Reductase Inhibitors

Jaroslaw Polanski *, Andrzej Bak, Rafal Gieleciak and Tomasz Magdziarz

Department of Organic Chemistry, Institute of Chemistry, University of Silesia, PL-40-006 Katowice Poland

* Author to whom correspondence should be addressed; e-mail: polanski@us.edu.pl

Received: 31 May 2004 / Accepted: 21 Oct 2004 / Published: 31 December 2004

Abstract: We have used SOM and grid 3D and 4D QSAR schemes for modeling the activity of a series of dihydrofolate reductase inhibitors. Careful analysis of the performance and external predictivities proves that this method can provide an efficient inhibition model.

Keywords: Self-organizing neural network, 3D QSAR, 4D QSAR, SOM-4D QSAR, CoMSA.

Introduction

Drug discovery is a complex issue that lacks a general approach. Drugs are mainly synthetic products developed by chemists. However, in this context *the most fundamental and lasting objective of synthesis is not a production of new compounds but the production of properties,* as commented by Hammond and cited by Sharpless and co-workers [1]. This fact clearly makes QSAR (Quantitative Structure-Activity Relationships), in its broadest sense, an essential and irreplaceable method in this field. However, more and more sophisticated tools are needed for the transformation of the molecular structure into the compound property space. Generally, the drug-receptor interactions are complex phenomena, which cannot be easily described. Therefore, a QSAR strategy of the comparison of a series of drug ligands separately from the receptor structure has evidently limitations. In 3D or 4D QSAR molecular superimposition that should be performed for the compound series can be mentioned here as an illustrative example. By performing superimposition, intentionally or by default, but generally independently from the receptor structure, we are assuming that molecular recognition proceeds *with exactly the same mechanism and in the same place within the receptor macromolecule.* The so-called similarity paradox (very similar molecules can evoke completely different biological activity) clearly proves that in reality this assumption is not true. Thus we need to make QSAR

insensitive, as much as possible, to the noise that may appear in the data. The application of robust modeling methods [2], i.e., such that they are resistant to uncertain data may be a key to success. Neural networks can be an example of such a technique that has been successfully used in drug design [3-5]. On the other hand, multivariate nonlinear regression has also been reported as an efficient alternative to neural techniques [6-8].

Our previous publications described possible applications of self-organizing neural networks for modeling 3D and 4D QSAR [9-18]. A self-organizing neural network is an unsupervised learning scheme consisting only of a single layer, usually two-dimensional rectangular or hexagonal grid of nodes (neurons). Different distance metrics can be used to define neighborhood relations between these neurons [19-22]. SOM (self-organizing map) network is designed to process multidimensional (N-dimensional) data vectors by distributing them between the neurons in such a way that similar inputs are put closer (into the neurons that are closer neighbors) to each other than those less similar. It is worth mentioning that the method preserves the topology of the processed input object.

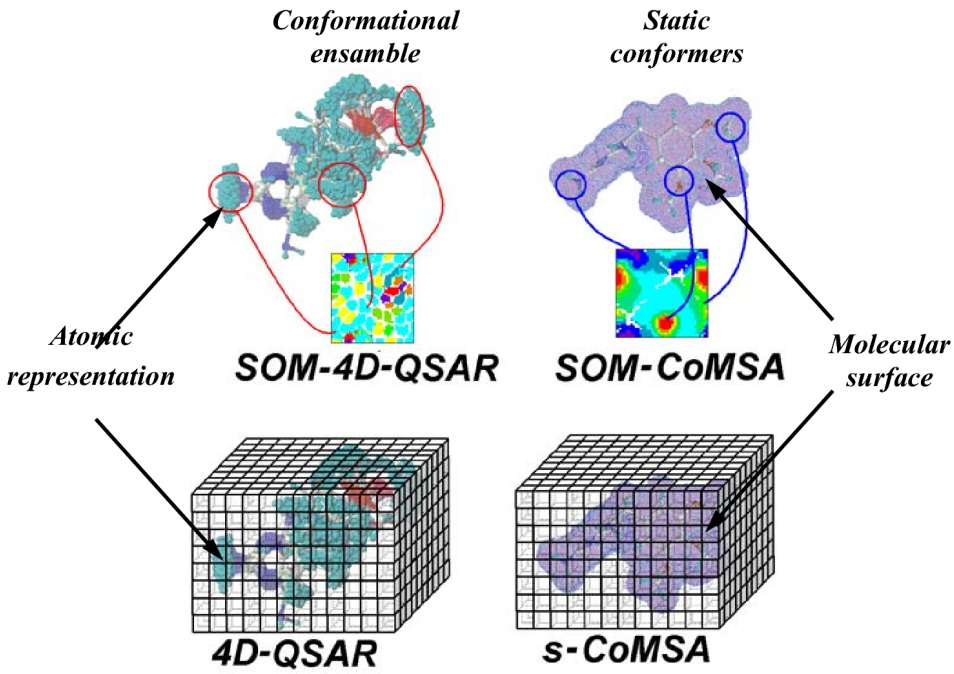
Considerable progress can be observed over the past decades in molecular development and design, in particular, in drug design. This includes new emerging disciplines and strategies that have appeared in this field. Combinatorial chemistry came as a first alternative to the traditional design and synthetic techniques. Genomics and related fields (e.g., chemo- and pharmacogenomics) have brought about an explosion of the data available for molecular design. The question may arise as to how much these new directions influence traditional methods. Do we still need traditional or multidimensional (3D or 4D) QSARs? Have traditional methods profited from these new directions? It can be clearly observed that generally, unnecessarily increasing the number of analyzed molecules, new methods investigate much larger *data populations*, irrespective of any technical problems encountered in such cases. 4D QSAR can be an example of such a technique. Basically, 4D QSAR investigates the conformational space of the molecular objects. However, in this calculation we generate for a single molecule the enormous number of conformers that investigates different spatial region. Actually, it is the likelihood of a formation of common 3D patterns of a series of molecules that is sought after by the molecular dynamics simulations. Many-fold replications of the molecules by different conformer representations allow for the increase of the chances for proper receptor structure mapping by the respective ligand structures. All this makes 4D QSAR scheme of the much more probabilistic nature, if compared to the 3D-QSAR.

In this publication we discuss the application of the SOM neural network for a QSAR scheme, in particular the SOM-4D-QSAR. Moreover, we compare this method with Comparative Molecular Surface Analysis (CoMSA) – a 3D QSAR method by the SOM neural network coupled with the Partial Least Squares Analysis (PLS) for a series of dihydrofolate reductase (DHFR) inhibitors [23].

Results and Discussion

Scheme 1 illustrates the methods used, i.e. SOM-CoMSA [9], s-CoMSA (sector – CoMSA) [24], grid-4D-QSAR and SOM-4D-QSAR [17].

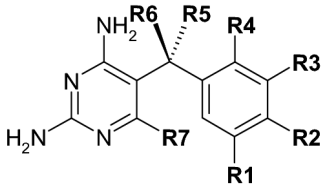
Scheme 1.



A series of dihydrofolate reductase inhibitors (DRI) are given in Table 1.

Hopfinger *et al.* analyzed this series in a publication that introduces a 4D QSAR method [23]. Thus we decided to use the same series for the comparison of the SOM versions of 3D and 4D QSAR methods. Several 3D techniques failed to model QSARs for these compounds; however; Hopfinger's 4D QSAR appeared to give a final regression equation (R^2=0.957, q^2=0.885, s=0.34) optimized by genetic algorithm (GA), performed after initial PLS. Instead of GA, we used in our method the PLS algorithm coupled with variable elimination. It is usually believed that variable elimination is not as important in PLS modeling as in standard regression procedure, because basically data transformed by PLS include this *part* of the original data that is essential for the description of the appropriate answer. However, data elimination can also be applied in PLS modeling, e.g. in Uninformative Variable Elimination (UVE) method developed by Centner *et al.* [25]. Compare references [7, 26] for the detailed investigations of variable selection in multiregression.

Table 1. The set of substituted 2,4-diamino-5-benzylpyrimidine inhibitors of *Escherichia coli* DHFR and their activity data [23].



No.	R1	R2	R3	R4	R5	R6	R7	log(1/I_{50})
1	OCH_3	OCH_3	OCH_3	H	H	H	H	8.23
2	OCH_3	OCH_3	OCH_3	CH_3	H	H	H	5.85
3*R*	OCH_3	OCH_3	OCH_3	H	OH	CH_3	H	4.00
4*S*	OCH_3	OCH_3	OCH_3	H	OH	CH_3	H	4.00
5	OCH_3	OCH_3	OCH_3	H	$=CH_2$		H	5.60
6*R*	OCH_3	OCH_3	OCH_3	H	H	CH_3	H	5.35
7*S*	OCH_3	OCH_3	OCH_3	H	H	CH_3	H	5.35
8	OCH_3	Br	OCH_3	H	H	H	H	8.53
9	OCH_3	OH	OCH_3	H	H	H	H	7.96
10	OCH_3	OH	OCH_3	H	H	H	CH_3	6.52
11	OCH_3	OCH_3	OCH_3	H	H	H	CH_3	7.00
12	OH	H	OH	H	H	H	H	2.78
13	H	H	H	H	H	H	H	5.71
14	CH_2OH	H	CH_3OH	H	H	H	H	5.83
15	H	H	Cl	H	H	H	H	6.14
16	H	Br	H	H	H	H	H	6.30
17	OCH_3	H	H	H	H	H	H	6.40
18	OCH_3	H	OCH_3	H	H	H	H	7.75
19	CH_3	H	CH_3	H	H	H	H	7.45
20	H	C_6H_5	H	H	H	H	H	6.40

In our previous publications we have shown that this method as well as its modifications, i.e., modified UVE (m-UVE) and iterative variable elimination (IVE) can be used in 3D QSAR schemes [11]. This allows identifications of the molecular areas important for the interactions with biological receptors or enzymes, so-called interaction pharmacophore elements (IPEs).

Figure 1. The 4D QSAR models of the DRI series, performed using the occupancy and charge type IPEs. The numbers indicate the q^2 and s performances, and optimal number of the PLS latent variables included in the model; after UVE and (modified procedure [11]) IVE data elimination, respectively

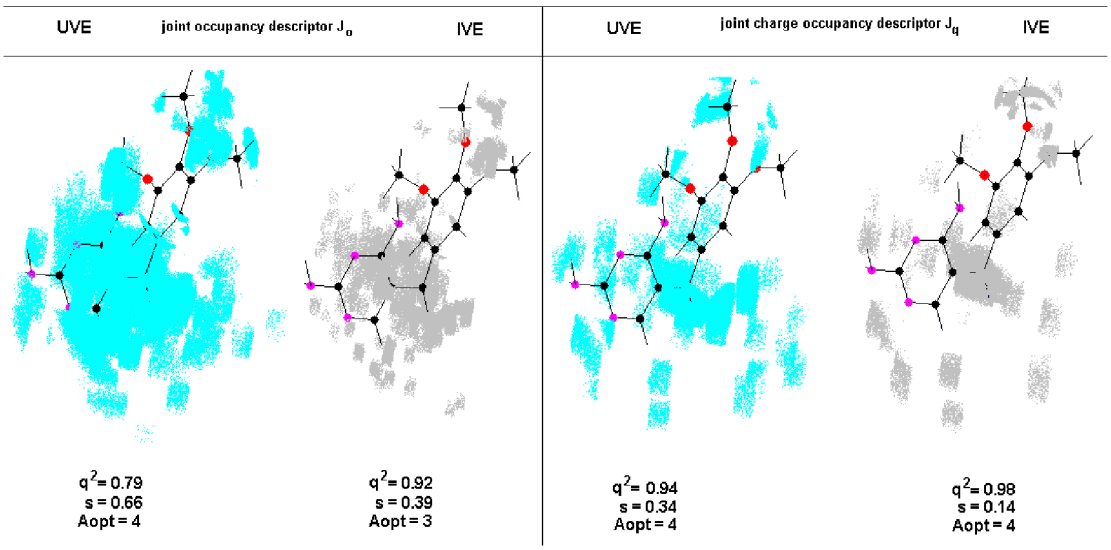
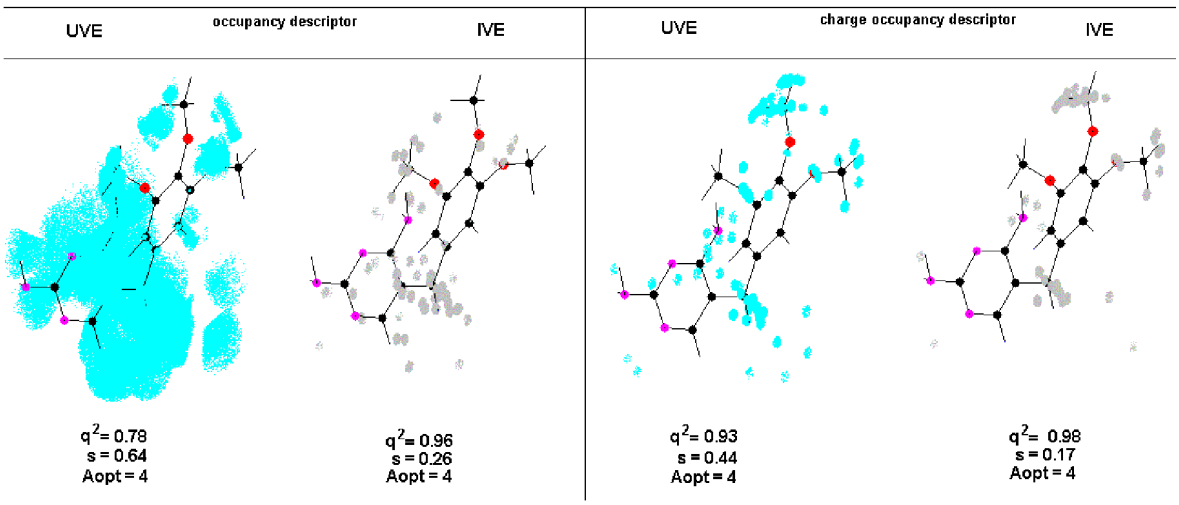


Figure 2. The SOM-4D QSAR models of the DRI series, performed using the occupancy and charge type IPEs. The numbers indicate the q^2 and s performances, and optimal number of the PLS latent variables included in the model; after UVE and IVE data elimination, respectively



The performance of the 4D QSAR PLS models obtained without data elimination ranges from q^2=0.30-0.43 and for the best model takes a value of q^2= 0.43, s=1.10, with 4 PLS components. After variable elimination these values can improve, as shown in Figures 1 and 2. This outperforms classical Hopfinger's 4D QSAR with GA. However, these data refer to the series without molecule **12**, which is an evident outlier according to our results. This can result from a fact that the activity of this compound evidently differs from the rest of the series. This may indicate some differences in the drug receptor interaction mechanism.

During modeling we always estimated an optimal number of the PLS components, but the maximal model complexity (a number of PLS components) was truncated not to exceed four. Our results only slightly depend upon the method used, i.e., classical grid method or its SOM version, and superposition mode. Figures 1 and 2 compare the IPEs revealed for DRI by SOM-4D-QSAR to that of 4D QSAR-PLS-UVE (IVE, m-UVE) methods. The performances have also been compared.

In the original publication Hopfinger *et al.* did not perform any additional model validation. However, according to the current knowledge, the q^2 value is not a sufficient criterion for verifying model quality. Thus, the description of the series using a few original variables (individual grid cells) without validation of the external predictions seems to be risky. Therefore, we divided the series into two groups of the training (compounds: 1-11 and 13-15) and test sets (compounds: 16-20) and verified model calculated for the training set by the residual error estimated for the values predicted in the test set. The best model was obtained for grid-4D-QSAR with joint occupancy type descriptors (I_c), which is characterized by q^2= 0.96, s=0.10 and standard deviation of error of prediction (SDEP) = 0.64 (with 10 PLS components. Of course a high number of PLS components makes a problem. On the other hand, this complexity is determined by model optimization. After data elimination that force lower complexity the performance is only slightly lower: q^2=0.96, s=0.12, SDEP = 0.61 with 4 PLS components. Thus, in this particular case lower complexity does not improve model predictability (SDEP value).

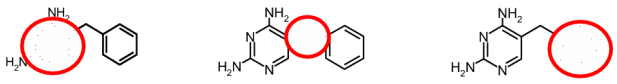
We think there are few interesting observations that appeared from the analysis of the results given in Figure 1. Thus, the molecular areas indicated by our analysis with the *occupancy type descriptors* are similar to those revealed in Hopfinger's work, in which *these type descriptors* were also used. The inclusion of *charge descriptors* improves model quality given by the q^2 and SDEP values. There is also some important regularity that can be observed during data elimination in the PLS model. In fact, we observed that for the models of the high predictivity data elimination cannot improve model quality. However, for the poor initial models, UVE (IVE) data elimination can bring an important improvement.

Both SOM- and grid- 4D-QSAR analyzed schemes provide comparable results. Table 2 compares performances of these schemes with 3D-QSAR modeling of the series activity. We used for this purpose the SOM and sector version of the Comparative Molecular Surface Analysis [24]. The results of 3D QSAR modeling are evidently worse than 4D QSAR. This indicates that conformational flexibility of the benzylpirymidine series makes 4D QSAR more efficient in modeling their inhibiting properties.

Conclusions

We used SOM and grid 3D and 4D QSAR schemes for modeling the activity of a series of dihydrofolate reductase inhibitors. We used PLS with UVE (IVE) for modeling all schemes. Careful analysis of the performances and external predictivities proves that this method can provide an efficient inhibition model.

Table 2. 3D QSAR results.



a ***b*** ***c***

		Superposition mode		
		a	b	c
CoMSA				
all [b]	MD[a]	0.5	0.5	0.5
	q^2	0.62	0.72	0.64
	S	1.17	1.01	1.08
Training/test set	MD	0.5	0.5	0.5
	q^2	0.59	0.64	0.71
	S	1.02	0.91	0.90
	SDEP	1.25	0.96	1.20
IVE	max A[c]	6	5	6
	MD	0.5	0.5	0.5
	q^2	0.62	0.87	0.71
	S	0.89	0.58	0.79
	SDEP	0.81	0.72	0.78
s-CoMSA				
all	sector size	1	1	3
	q^2	0.38	0.56	0.47
	s	1.50	0.90	1.01
Training/test set	sector size	1	1	1
	q^2	0.54	0.70	0.69
	s	1.96	1.59	1.26
	SDEP	1.42	1.32	1.42
IVE	sector size	1	1	4
	max A	4	1	3
	q^2	0.73	0.59	0.68
	s	0.83	0.92	0.85
	SDEP	0.93	1.15	1.10

[a] MD – Maximal distance for comparative Kohonen maps [10] [b] Models without variable elimination [c] Maximal number of PLS components during variable elimination procedure (IVE-PLS) [11].

Acknowledgments

The authors thank Professor Johann Gasteiger of the University of Erlangen-Nürnberg, BRD both for his valuable discussions and for facilitating access to the CORINA, PETRA, SURFACE, and KMAP programs. The financial support of the KBN Warsaw under grants no: KBN – 4T09A 088 25 3T09A 01127 and PBZ 040 P04/08 is gratefully acknowledged. RG thanks Foundation for Polish Science for an individual grant.

Experimental

The SOM-CoMSA [9], s-CoMSA [24], grid-4D-QSAR [23] and SOM-4D-QSAR [17] procedures were described in the cited previous publications, respectively.

Model builders

All the experimental data, i.e. biological activities for the dihydrofolate reductase inhibitors were extracted from ref. [23] and are given in Table 1

Kohonen mapping

The competitive Kohonen strategy [19] was used to construct a two-dimensional topographic map obtaining the signals from the points sampled randomly at the molecular surface. As molecular surfaces are continuous the plane of projection was also selected to be a continuous surface. Thus we used a torus for this purpose, which was cut along two perpendicular lines and then spread into a plane. Each neuron, j, was then defined by three weights, w_{ji}. The competitive training of the network was based on the rule that each point, s, of the molecular surface was projected into that neuron, sc, that has weights, w_{ci}, that come closest to the Cartesian coordinates, x_{si}, of this point, s, (eq. 1).

$$out_{sc} \leftarrow \min\left[\sum_{i=1}^{m}\left(x_{si} - w_{ji}\right)^2\right] \quad (1)$$

A projection of the molecular electrostatic potential (MEP) value from the surface points, s, into such a two-dimensional arrangement of neurons, after calculating the average MEP value within this particular neuron and scaling this values into the respective colors results in the so called feature map.

Comparative Kohonen mapping

In fact, such a map illustrates the property (MEP) of a single molecule. As however, the weights of the Kohonen network contain the shape of the certain molecular surface, it can be used to compare the geometries of molecular surfaces of other molecules. In such a method the trained Kohonen network is processing the signals coming from the surface of other molecule(s), i.e., the electrostatic potential of each input vector was projected through the network to obtain a series of comparative maps both for the template molecule and each analyzed molecule. The respective electrostatic potential values from the surfaces of the processed molecules were then projected into such a network allowing us to compare these parts of the molecule surfaces that can be superimposed. If the surfaces cannot be

superimposed on the reference molecule (template) then the respective output neurons get no signal from the molecules processed.

All the molecules were superimposed before the calculation of molecular surfaces. The superimposition was performed as shown in Table 2. In practice, we used Match3D program [27] for performing this operation. The KMAP 3.0 program [27] was used for the simulation of Kohonen networks. The size of the Kohonen networks amounts from *10x10* to *30*x*30* neurons. The output of this program was used for the calculation of the mean electrostatic potential values within each neuron and respective feature maps were transformed to a respective 10^2, 20^2 and 30^2 element vectors.

4D QSAR calculation

We used Hopfinger's spatial grid system [23] for coding molecules. The molecules after AM1 (Austin model 1) optimization were used as initial structures in the molecular dynamic simulation (MDs). Each 3D structure is the starting point in generating conformational ensemble profile (CEP). Molecular dynamics was performed using the Sybyl software [28] with standard Tripos force field. 2500 conformations were sampled for each analogue. Partial atomic charges were calculated using the semiempirical AM1 Hamiltonian (HYPERCHEM package [29]). The alignment of the molecules was the next step of the 4D-QSAR analysis. We aligned the molecules according to the previous rules of the Hopfingers' study [23]. Individual conformers are placed in the grid cell space surrounding the aligned compounds. We applied cubic grid lattice of 20 Å on each side with grid cell resolution of 1, 2 or 0.5 Å, respectively. Different types of grid cell occupancy descriptors (GCODs) were considered and calculated for the indicated atoms referred to as interaction pharmacophore elements (IPE). Apart from, the GCODs used by Hopfinger et al. [23], we applied in our current work the absolute charge occupancy (A_q) for the chosen IPE atoms of compound c defined as

$$A_q(c,i,j,k,N) = \sum_{t=0}^{T} O_t(c,i,j,k) \times q / m \tag{2}$$

where m means the number of the atoms of compounds, c present in the cell (i,j,k) at time t, q means the sum of partial atoms of charges present in some cell at time t, T is the length of the time in MDs. N is the number of sampling MDs steps. The joint (J_q) and self charge occupacy (S_q) with the most active reference compound R were defined after following equations:

$$J_q(c,i,j,k,N) = \sum_{t=0}^{T} O_t(c,i,j,k) \cap O_t(R_q,i,j,k) \times q / m \tag{3}$$

$$S_q(c,R,i,j,k,N) = \sum_{t=0}^{T} \{O_t(c,i,j,k) - [\sum_{t=0}^{T} O_t(c,i,j,k) \cap O_t(R,i,j,k)]\} \times q / m \tag{4}$$

We used the MATLAB [30] environment to program the calculation of the above mentioned descriptors. The Partial Least Squares (PLS) method with variable elimination was used to estimate the relationship between independent variables (GCODs) and corticosteroid binding globulin (CBG) affinity.

Calculation of the molecular surface (s-COMSA) descriptors based on virtual cubic grid

For the calculation of shape descriptors we applied formalism similar to Hopfinger's 4D-QSAR grid coding system using the absolute type descriptors, as given by the above mentioned equations. However, unlike in 4D QSAR our method compares single conformers. Thus, each 3D molecular representation is placed in its own virtual cubic grid and molecular surface is calculated, respectively. The electrostatic potential is calculated for the points randomly sampled on the molecular surface and a mean value of the electrostatic potential corresponding to the respective points found in each grid cell is used to describe this cell. Grid cells are unfolded into vectors and vectors describing all molecules of the series are aligned into a matrix. Grid cells that are empty for all molecules in the series analyzed are eliminated and the resulted matrix was used for further calculations using the PLS method.

PLS analysis

Obtained vectors were processed by the PLS analysis with a leave-one-out cross-validation procedure. The PLS procedures were programmed within the MATLAB environment (MATLAB) [30].

A PLS model was constructed for the centered data and its complexity was estimated on the basis of the leave-one-out cross-validation procedure (CV). In the leave-one-out CV one repeats the calibration m times, each time treating the i-th left-out object as the prediction object. The dependent variable for each left-out object is calculated on the basis of the model with one, two, three etc. factors. The Root Mean Square Error of CV for the model with j factors is defined as:

$$\mathrm{RMSECV_j} = \sqrt{\frac{\sum_i (\mathrm{obs_i} - \mathrm{pred_{i,j}})^2}{\mathrm{m}}} \tag{5}$$

where obs denotes the assayed value; pred - predicted value of dependent variable and i refers to the object index, which ranges from 1 to m. Model with k factors, for which RMSECV reaches a minimum, is considered as an optimal one.

We used the performance metrics that are accepted and widely used in CoMFA analyses, i.e., cross-validated q_{cv}^2

$$q_{CV}^2 = 1 - \frac{\sum (obs_i - pred_i)^2}{\sum (obs_i - mean(obs))^2} \tag{6}$$

where *obs* - the assayed values; *pred* - predicted values, *mean* - mean value of *obs* and *i* refers to the object index, which ranges from 1 to *m*; and cross-validated standard error *s*

$$s = \sqrt{\frac{\sum (obs_i - pred_i)^2}{m - k - 1}} \tag{7}$$

where *m*- number of objects, *k*- number of the PLS factors in the model.

Before the PLS analysis was performed the descriptors were centered and this operation was repeated for each cross-validation run.

The quality of external predictions was measured by the Standard Deviation of Error of Prediction (SDEP) parameter:

$$SDEP = \sqrt{\frac{\sum_i (pred_i - obs_i)^2}{n}} \quad (8)$$

where *pred* – predicted value, *obs* – observed value.

References and Notes

1. Kolb, H.C.; Finn, M.G.; Sharpless, K.B. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angew. Chem., Int. Ed. Engl.* **2001**, *40*, 2004-2021.
2. Buden, F.R.; Winkler, D.A. Robust QSAR models using Bayesian regularized neural networks. *J. Med. Chem.* **1999**, *42*, 3183-3187.
3. Schneider, G.; Wrede, P. Artificial neural networks for computer-based molecular design. *Prog. Biophys. Mol. Biol.* **1998**, *70*, 175-222
4. Polanski, J. Molecular shape analysis. In: *Handbook of Chemoinformatics*; Gasteiger J. (ed.); Wiley-VCH Verlag: Weinheim, **2003**; pp. 302-319.
5. Anzali, S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J.; Teckentrup, A.; Wagener M. The use of self-organizing neural networks in drug design. *Perspect. Drug Discov. Design* **1998**, *9/10/11*, 273-299.
6. Lucic, B.; Trinajstic, N. Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121-132.
7. Lucic, B.; Trinajstic, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610-621.
8. Lucic, B.; Amic, D.; Trinajstic, N. Nonlinear Multivariate Regression Outperforms Several Concisely Designed Neural Networks in QSPR Modeling. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 403-413
9. Polanski, J.; Walczak, B. The comparative molecular surface analysis (CoMSA): a novel tool for molecular design. *Comp. Chem.* **2000**, *24*, 615- 625.
10. Polanski, J.; Gieleciak, R.; Bak, A. The comparative molecular surface analysis (CoMSA) – a nongrid 3D QSAR method by a coupled neural network and PLS system: Predicting pK_a values of benzoic and alkanoic acids. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 184-191.
11. Polanski, J.; Gieleciak, R. The comparative molecular surface analysis (CoMSA) with modified uninformative variable elimination-PLS (UVE-PLS) method: application to the steroids binding the aromatase enzym. *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 656-666.
12. Polanski, J.; Gieleciak, R.; Wyszomirski, M. Comparative molecular surface analysis (CoMSA) for modeling dye-fiber affinities of the azo and antraquinone dyes, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1754-1762.
13. Polanski, J.; Gieleciak, R.; Wyszomirski, M. Mapping dye pharmacophores by the comparative molecular surface analysis (CoMSA): application to heterocyclic monoazo dyes. *Dyes Pigm.* **2004**, *62*, 63-78.

14. Polanski, J.; Gasteiger, J.; Jarzembek, K. Self - Organizing neural networks for screening and development of novel artificial sweetener candidates. *Combin. Chem. High Throughput Screen.* **2000**, *3*, 481-495.
15. Polanski, J.; Gieleciak, R. Comparative molecular surface analysis: a novel tool for drug design and molecular diversity studies, *Mol. Diversity* **2003**, *7*, 45-59.
16. Polanski, J. Self-organizing neural networks for pharmacofore mapping. *Adv. Drug Deliv. Rev.* **2003**, *55*, 1149-1162.
17. Polanski, J.; Bak, A. Modeling steric and electronic effects in 3D and 4D-QSAR schemes: Predicting benzoic pK_a values and steroid CBG binding affinities. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2081-2092.
18. Polanski, J.; Gieleciak, R.; Bak, A. Probability issues in molecular design: Predictive and modeling ability in 3D-QSAR schemes. *Comb. Chem. High T. Scr.* In press
19. Kohonen, T. *Self-Organization and Associative Memory*, 3rd Edition, Springer Verlag: Berlin, **1989**.
20. Zupan, J.; Gesteiger, J. *Neural Networks in Chemistry and Drug Design, 2nd Edition*; Wiley–VCH: Weinheim, **1999**.
21. Melssen, W.J.; Smits, J.R.M.; Buydens, L.M.C.; Kateman, G. Tutorial: Using artificial neural networks for solving chemical problems. Part II. Kohonen self-organising feature maps and Hopfield networks, *Chemometer. Intell. Lab. Syst.* **1994**, *23*, 267-291.
22. Kohonen, T. The Self-Organizing Map (SOM), http://www.cis.hut.fi/projects/somtoolb.shtml.
23. Hopfinger, A.J.; Wang, S.; Tokarski, J.S.; Jin, B.; Albuquerque, M.; Madhav, P.J.; Duraiswami, C. Construction of 3D QSAR models using the 4D QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509-10524.
24. Polanski, J.; Gieleciak, R.; Magdziarz, T. The grid formalism for the comparative molecular surface analysis: application to the CoMFA benchmark steroids, azo dyes and HEPT derivatives. *J. Chem. Inf. Comput. Sci.* In press
25. Centner, V.; Massart, D. L.; de Noord, O.E.; de Jong, S.; Vandeginste, B. M. V.; Sterna, C. Elimination of uninformative variables for multivariate calibration. *Anal. Chim. Acta.* **1996**, *330,* 1-17.
26. Pilizota, T.; Lucic, B.; Trinajstic, N. Use of variable selection in modeling the secondary structural content of proteins from their composition of amino acid residues *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 113-121
27. Gasteiger, J. Match3D; KMAP for the information see: http://www2.ccc.uni-erlangen.de.
28. Sybyl 6.5. program, available from the Tripos Inc., St. Louis, MO, USA: http://www.tripos.com.
29. *HyperChem 5.0*, available from HyperCube Inc., Gainesville, FL, USA: http://www.hyper.com.
30. *MATLAB 6.5*, available from The Mathworks Inc., Natick, MA, USA, http://www.mathworks.com.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry
Bioorganic & Medicinal Chemistry 14 (2006) 1630–1643

3D QSAR study of hypolipidemic asarones by comparative molecular surface analysis

Tomasz Magdziarz,[a] Bozena Łozowicka,[b] Rafał Gieleciak,[a] Andrzej Bąk,[a] Jarosław Polański[a,*] and Zdzisław Chilmonczyk[b,c]

[a]*Department of Organic Chemistry, Institute of Chemistry, University of Silesia, PL-40-006 Katowice, Poland*
[b]*Department of Organic Chemistry, Institute of Chemistry, University of Bialystok, J. Pilsudskiego 11/4, PL-15-443 Bialystok, Poland*
[c]*National Institute of Public Health, Chełmska 30/34, 00-725 Warszawa, Poland*

Received 3 August 2005; revised 30 September 2005; accepted 6 October 2005
Available online 3 November 2005

Abstract—Three-dimensional quantitative structure–activity relationship (3D QSAR) modeled for α-asarone derivatives using the comparative molecular surface analysis (CoMSA) allowed us to reveal a correlation between the activity of these compounds and the electrostatic potential at the molecular surface. The grid formalism (s-CoMSA) allowed us to indicate a pharmacophore that is of key importance for compound activity. The CoMSA formalism coupled with the iterative variable elimination method gives a highly predictive model.

1. Introduction

Atherosclerosis and subsequent cardiovascular diseases still remain one of the major death risks in the industrialized and developing countries. The initiation of atherosclerosis is most likely caused by inflammatory responses, hyperlipidemia, and blood clotting factors. Many factors were found to be associated with the development of atherosclerosis and cardiovascular diseases.[1–5] The link between elevated cholesterol and coronary heart diseases (CHD) has been clearly established, and clinical trials have found that a 1% reduction in serum total cholesterol reduced CHD risk by 2%.[6]

Basically, cells (except for hepatic and ileum cells) do not synthesize cholesterol de novo but obtain it from blood and the cholesterol that accumulates in atherosclerotic lesions originates, primarily, in plasma lipoproteins.[7] The lowering of abnormally elevated levels of atherogenic lipoproteins (chylomicrons, very low density lipoproteins—VLDL-C, and low density lipoproteins—LDL-C) is now accepted as the first line of approach to the treatment of hyperlipidemic patients.[8]

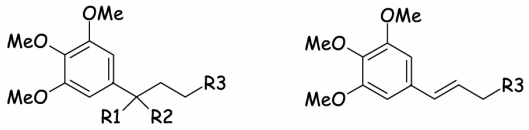
Increasing attention is also being focused on other lipoprotein fractions, such as high-density lipoprotein cholesterol (HDL-C) and triglycerides (TG), as additional potential targets of therapy. Elevated serum TG combined with low HDL-C, a condition often associated with smaller, dense LDL particles, is frequently referred to as atherogenic dyslipidemia or the 'lipid triad.'[9]

At present, the therapy of lipid metabolism disorders is based on the use of drugs having hypolipidemic activity like fibric acid derivatives, anion exchange resins which sequester the bile acids, the inhibitors of 3-hydroxymethyl-coenzyme A (3-HMG Co-A) reductase, en enzyme involved in de novo sterol synthesis, probucol, lifibrol, and many others.[10,11] α-Asarone (**1**) is an active component of *Acorus calamus* Linn, *Acorus gramineus* Soland, and *Guatteria gaumeri*.[12] α-Asarone and its analogues are known to be endowed with hypolipidemic activity[13,14] and have been the subject of many pharmacological,[15–17] toxicological,[18,19] synthetic,[20] and QSAR studies.[21,22]

Different α-asarone structural features were found to influence the hypolipidemic activity. Chamorro et al. examined analogues possessing a dimethoxylated unconjugated propenyl side chain (mice) and Cruz et al.[23] analogues with saturated side chain (rats). Analysis of a length of a hydrophobic chain has shown that the most active analogues had the shortest side chain.

Keywords: α-Asarone analogues; Hypolipidemic activity; QSAR study.
* Corresponding author. Tel.: +48 32 3591197; fax: +48 32 2599978; e-mail addresses: bozena@uwb.edu.pl; polanski@us.edu.pl; chilmon@il.waw.pl

0968-0896/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2005.10.014



R_1, R_2=O; R_1=H, R_2=OH; R_3=amine

Figure 1. Molecular formulae of compounds **7**–**13**.

During our work on α-asarone analogues, we synthesized several new compounds and examined their hypolipidemic activity.[21,24–26] We also examined a relationship between the hypolipidemic activity and molecular features. It appeared that the asarone activity could be described, at least in part, using the pseudo- or mini-receptor models.[21,27] However, we could not have indicated any clear molecular rule controlling the activity of these compounds.[24–26]

Some data concerning the compounds **1**–**40** were preliminarly reported[27] however, no experimental details are available for compounds **7**–**13**. Thus, in the present publication we describe the synthesis and biological evaluation of compounds **7**–**13** (Fig. 1, Scheme 1).

We concentrate on the understanding of the structural basis of the compounds' pharmacological activity. Since our previous models did not completely explain the structure–activity relationship, we investigated 3D QSAR by comparative molecular surface analysis, a novel method described in our previous publications.[28–36] We have found that this method indicates a clear molecular basis for the activity.

2. Methods

2.1. Data sets for the analysis

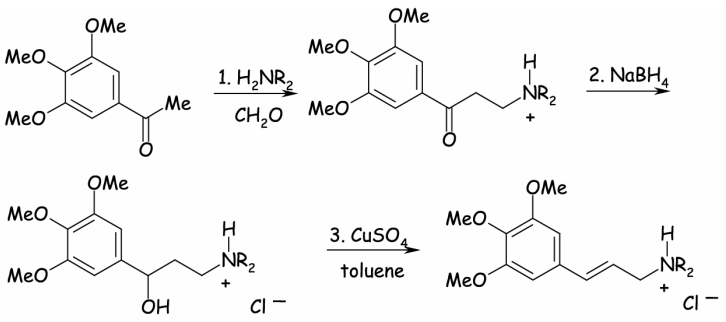
The chemical structures of the α-asarone derivatives are shown in Table 1. Hypolipidemic activity reported was measured (see Section 2.2, Table 1). As the activity data in all our QSAR calculation, we used the atherogenic index $I_{TG/HDL}$. It is calculated by the ratio of the triglyceride concentration (TG) [mmol/L] to the HDL cholesterol concentration (HDL-C) [mmol/L]. The atherogenic index for the analyzed series ranges from 0.10 (high activity) to 3.28 (low activity) (Table 2).

2.2. Hypolipidemic activity

Wistar rats weighing 300–320 g male were bought from Nofer Institute of Occupational Medicine, Łódź, Poland. The animals were equally divided into groups of five animals each and maintained at temperature of 22 ± 2 °C, 45–80% relative humidity, and every 12 h periods changing of light and darkness. All rats were fed a high cholesterol diet (Murigran enriched with cholesterol 1%, sodium cholate 0.2%, and olive oil 5%) for 7 days (Murigran-Łomna near Warsaw). Rats fed with laboratory chow for the same duration as above were used as noncholesterol control group. Compounds, diluted in oil, were administered through gastric intubation at 80 mg/kg once a day for the duration of the experiment. Group receiving clofibrate (150 mg/kg) served as positive control. Compounds, diluted in oil, were adjusted so that the rats were administered a volume of 5 mL/kg of body weight. Rats in the control group received a similar volume of vehicle. At the end of seven-day period, each animal was fasted for 16 h and anesthetized. Blood samples were collected through ocular puncture and centrifuged at 3000 rpm. Total cholesterol (TC), HDL-cholesterol (HDL-C), and triglycerides (TG) were determined using Alpha Diagnostics kits on Clinic System—700 Beckman. All the data were statistically analyzed by Student's *t* test. All the results were calculated against noncholesterol and cholesterol control groups.

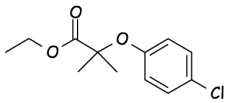
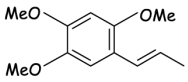
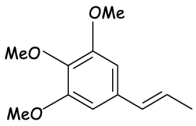
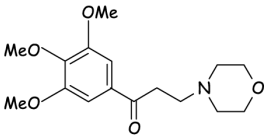
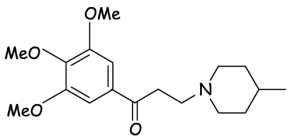
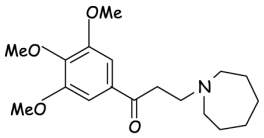
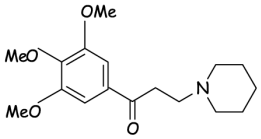
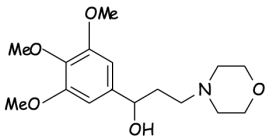
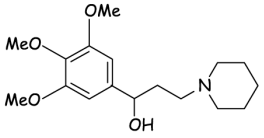
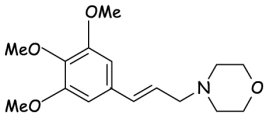
2.3. CoMFA analysis and molecular alignment

All modeling work was performed using the Sybyl 6.2 software package run on Silicon Graphics O2 workstation. The initial geometry was optimized using the standard Tripos force field (POWELL method) with 0.005 kcal/mol energy gradient convergence criterion and a distant-dependent dielectric constant. Charges were calculated using the Gasteiger–Marsilli method implemented in Sybyl. We used the FIT option of the Sybyl to align the compounds analyzed. Parent α-asarone, that is, a common fragment for all molecules, was chosen as a template. Alignment was carried out



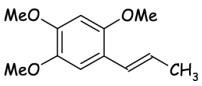
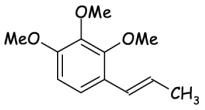
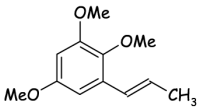
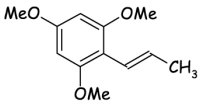
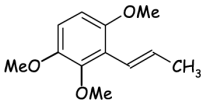
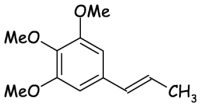
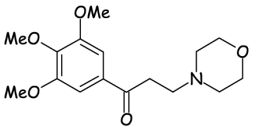
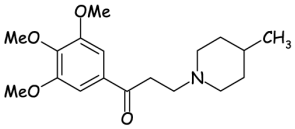
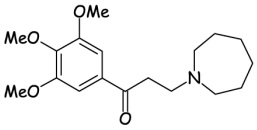
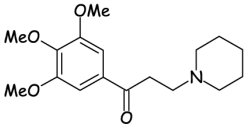
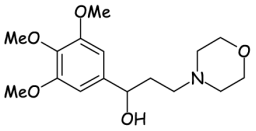
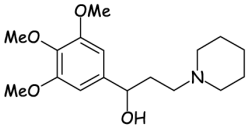
Scheme 1. General procedure synthesis of compounds **7**–**13**.

Table 1. Structures of α-asarone analogues and hypolipidemic activity[a]

Entry	Structure	TC	HDL-C	LDL-C	TG
Nonchole sterol diet **7**–**13**	...	−61 ± 0.16**	+152 ± 0.05**	−91 ± 0.07*	−39 ± 0.02**
Cholesterol diet (CD) **7**–**13**	...	100 ± 0.06[b]	100 ± 0.06[c]	100 ± 0.33[d]	100 ± 0.29[e]
CLO + CD		−23 ± 0.14**	+24 ± 0.06**	+7 ± 0.14	−27 ± 0.12**
1 + CD		+2.3 ± 0.24	+57 ± 0.06**	−43 ± 0.17**	+75 ± 0.30**
6 + CD		−16 ± 0.28	+56 ± 0.06**	−54 ± 0.14**	+8 ± 0.16
7 + CD		+7 ± 1.16	+71.0 ± 0.16*	−7.4 ± 1.13	+54 ± 0.31
8 + CD		−10 ± 0.84**	+35 ± 0.07**	−15.6 ± 0.88	−16 ± 0.19**
9 + CD		+18 ± 0.52**	+68 ± 0.09*	+12 ± 0.64	+12 ± 0.09**
10 + CD		+6 ± 0.58**	+65 ± 0.11*	−5 ± 0.50**	+26 ± 0.18
11 + CD		0 ± 1.59	+36 ± 0.09	−2 ± 1.62*	−23 ± 0.18
12		+13 ± 0.25*	+10 ± 0.06**	+20 ± 0.28*	−15 ± 0.10
13		+13 ± 0.38	+26 ± 0.09	+12 ± 0.54	+16 ± 0.33

[a] Expressed as a percentage of the cholesterol diet group (mean ± SD), $n = 6$.
[b] Cholesterol diet group 3.07 mmol/L.
[c] Cholesterol diet group 0.31 mmol/L.
[d] Cholesterol diet group 2.44 mmol/L.
[e] Cholesterol diet group 0.69 mmol/L; for compounds **7**–**13**.
* Significantly different from the result for the cholesterol diet group control at $p < 0.05$.
** Significantly different from the result for the cholesterol diet group control at $p < 0.01$.

Table 2. α-Asarone analogues and atherogenic index—$I_{TG/HDL}$ data

Compound	Molecular formulae	$I_{TG/HDL}$	TG (mmol/L)	HDL-C (mmol/L)
1		2.10	2.48	1.18
2		1.27	1.41	1.11
3		1.56	1.58	1.01
4		1.34	1.31	0.98
5		2.35	1.83	0.78
6		1.31	1.53	1.17
7		2.00	1.06	0.53
8		1.38	0.58	0.42
9		1.48	0.77	0.52
10		1.71	0.87	0.51
11		1.26	0.53	0.42
12		1.74	0.59	0.34

(*continued on next page*)

Table 2 (*continued*)

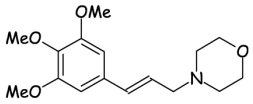
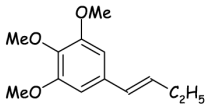
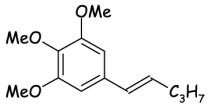
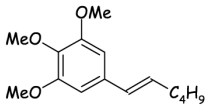
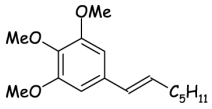
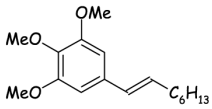
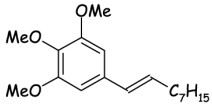
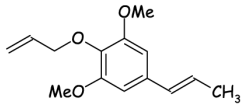
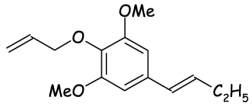
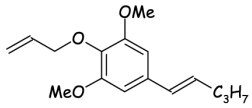
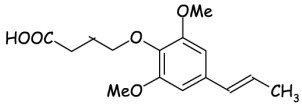
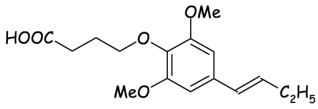
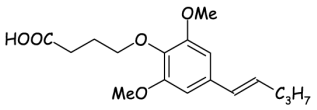
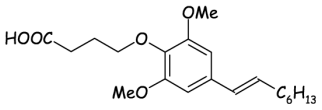
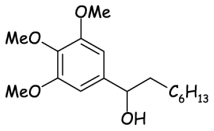
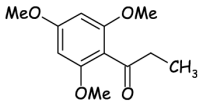
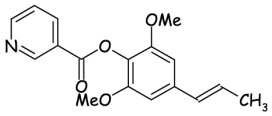
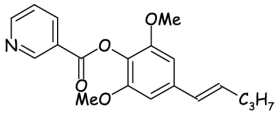
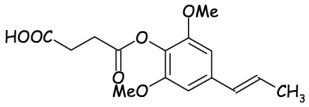
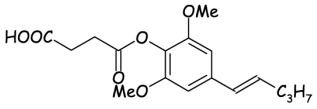
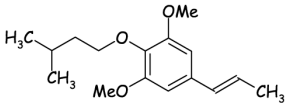
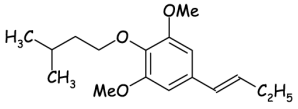
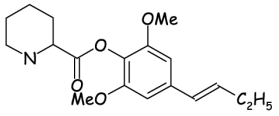
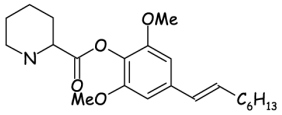
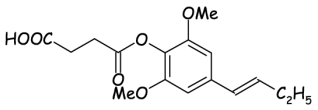
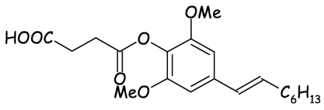
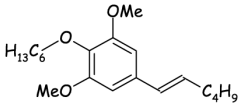
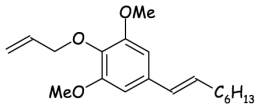
Compound	Molecular formulae	$I_{TG/HDL}$	TG (mmol/L)	HDL-C (mmol/L)
13		2.05	0.8	0.39
14		0.36	0.21	0.59
15		0.45	0.24	0.53
16		0.16	0.09	0.57
17		0.27	0.19	0.71
18		0.13	0.07	0.56
19		0.14	0.08	0.57
20		0.47	0.2	0.43
21		0.15	0.1	0.66
22		0.10	0.07	0.73
23		0.45	0.29	0.64
24		0.22	0.14	0.64

Table 2 (*continued*)

Compound	Molecular formulae	$I_{TG/HDL}$	TG (mmol/L)	HDL-C (mmol/L)
25		0.13	0.09	0.69
26		0.85	0.4	0.47
27		0.77	0.33	0.43
28		1.90	1.71	0.90
29		3.15	1.04	0.33
30		1.52	0.67	0.44
31		1.79	0.61	0.34
32		3.28	0.95	0.29
33		0.45	0.21	0.47
34		0.56	0.18	0.32
35		1.98	0.99	0.5
36		2.58	1.01	0.4

(*continued on next page*)

Table 2 (*continued*)

Compound	Molecular formulae	$I_{TG/HDL}$	TG (mmol/L)	HDL-C (mmol/L)
37		1.92	0.96	0.5
38		1.96	1.06	0.54
39		1.28	0.77	0.6
40		0.80	0.43	0.54

by superimposing the atoms of benzene ring. We used a single conformation for each molecule, which provided an alignment illustrated in Supplementary Materials Figure 1. The steric (Lennard-Jones) and electrostatic fields around the set of compounds were sampled with the probe atoms: sp^3 carbon (charge +1 and 0) and hydrogen (charge +1), on the rectangular grid that encompasses all aligned molecules (with margin of 3.0–4.0 Å). For each molecule the energies of 1769 grid points were calculated with 2 Å spacing in a lattice of $16 \times 10 \times 12$. We kept a convention to truncate the steric and electrostatic values at the level of 30.0 kcal/mol.

2.4. CoMSA analyses

For the calculation of shape descriptors we applied both grid (s-CoMSA) and neural formalisms described in our previous publications.[28–36] Thus, each 3D molecular representation is placed in its own virtual cubic grid and molecular surface is calculated, respectively. The electrostatic potential is calculated for the points randomly sampled on the molecular surface and a mean value of the electrostatic potential corresponding to the respective points found in each grid cell is used to describe this cell. Grid cells are unfolded into vectors and vectors describing all molecules of the series are aligned into a matrix. Grid cells that are empty for all molecules in the series analyzed are eliminated and the resulted matrix was used for further calculations using the PLS method. Alternatively, a CoMSA version with Kohonen self-organizing neural network (SOM-CoMSA) was used for comparison.

2.5. PLS analysis

The obtained vectors were processed by the PLS analysis with a leave-one-out cross-validation procedure. The PLS procedures were programmed within the MATLAB environment (MATLAB).[37]

A PLS model was constructed for the centered data and its complexity was estimated on the basis of the leave-one-out cross-validation procedure (CV). In the leave-one-out CV, one repeats the calibration m times, each time treating the ith left-out object as the prediction object. The dependent variable for each left-out object is calculated on the basis of the model with one, two, three, etc., factors.

The root mean square error of CV for the model with j factors is defined as

$$\mathrm{RMSECV} = \sqrt{\frac{\sum(\mathrm{obsd} - \mathrm{pred})^2}{m}},$$

where obsd denotes the assayed value; pred is the predicted value of dependent variable, which ranges from 1 to m. Model with k factors, for which RMSECV reaches a minimum, is considered as an optimal one.

We used the performance metrics that are accepted and widely used in CoMFA analyses, that is, cross-validated q^2_{cv}, s, RMS, and SDEP.

2.6. Iterative variable elimination

In our previous publications, we have shown that uninformative variable elimination (UVE)[38] as well as its modifications, that is, modified UVE (m-UVE) and iterative variable elimination (IVE), can be used in 3D and 4D QSAR schemes.[30,36,39] This enables the identifications of the molecular areas important for the interactions with biological receptors or enzymes. In the current calculation, we used iterative variable elimination (IVE-PLS)[30] that is a modification of the UVE

algorithm based on the analysis of the regression coefficients calculated by the PLS method. PLS allows presenting the relation between the Y answer and X predictors in a form of

$$Y = Xb + e,$$

where b is a vector of the regression coefficients and e the vector of the errors.

Thus, the UVE algorithm analyzes the reliability of the mean(b)/s(b) ratio (where s(b) means standard deviation of b). Then, only the variables of the 'relative' high mean(b)/s(b) ratio are included into the final PLS model. Instead of a single-step UVE procedure we used here an iterative algorithm based on the abs(mean(b)/s(b)) criterion to find the variables to be eliminated. This procedure includes:

1. Standard PLS analysis applied to analyze the matrices yielded from the s-CoMSA procedure with the leave-one-out cross-validation to estimate the performance of the PLS model (q^2_{cv}),
2. The elimination of the matrix column of the lowest abs(mean(b)/s(b)) value,
3. Standard PLS analysis of the new matrix without the column eliminated in step 2,
4. Iterative repetition of steps 1–3 to maximize the LOO q^2_{cv} parameter.

All procedures were programmed within the MATLAB environment (MATLAB).

3. Chemistry

The compounds employed in this study **7**–**13** (Fig. 1) were synthesized according to Scheme 1. The compounds were prepared by a modified method of Blike and Burckhalter.[40] Thus, an appropriately substituted amine was reacted with three equivalents of paraformaldehyde in absolute ethanol with the addition of equivalent of concentrated HCl to form an intermediate imminium salt, which was then reacted with the substituted trimethoxyphenone to afford the desired product (Scheme 1). This reaction works very well for small amines. Although the compounds were obtained in good yields, the amount of product sharply decreases as the size of the amine becomes bigger. Usually, the reaction time ranges from 2 to 4 h; however; it needs to be increased with the increasing amine bulkiness. The best results were obtained using ethanol/HCl solution and the products formed can be efficiently purified by recrystallization from ethanol/acetone (9:1).

4. Results and discussion

The hypolipidemic activity of compounds **7**–**13** was tested on male rats fed with cholesterol diet against clofibrate as a reference drug. Compounds **7**–**13** exhibited diversified hypolipidemic activity in rats. The most active compound **8** elevated the HDL-C level by 35% and diminished TC, LDL-C, and TG levels by 10%, 15.6%, and 16%, respectively (as compared to clofibrate with +24%, −23%, +7%, and −27% respective levels). Moreover, compounds **11** and **12** elevated HDL-C while diminishing TG levels.

The description of the molecular shape by spatial sectors was originally proposed by Purcel and Testa, and further improved by Motoc.[41] In this method, a molecule is separated into partitions of the spatial regions either filled or unfilled by atoms or groups of atoms of certain volumes. Using a similar idea but with improved formalism, we have developed the CoMSA method and proved that it can be a powerful tool for 3D QSAR modeling. Hasegawa reported some further CoMSA modifications.[42–44]

The statistical results of several CoMSA analyses are summarized in Table 3. For the comparison, we performed also standard CoMFA calculations. Cross-validated q^2_{cv} values obtained using these methods range from 0.58 (CoMFA) to $q^2_{cv} = 0.60$ (SOM-CoMSA). This proves a correlation between descriptors and the hypolipidemic activity of the asarone series.

CoMFA calculations provide the best results for the H^+ atomic probe, which points that electrostatic interactions limit compound activity (Table 3, model 1). All steric fields provided unpredictive models (the best model can be obtained for CH_3(0) probe; $q^2_{cv} = 0.35$).

All further results are reported for the s-CoMSA method. Standard method provides a relatively low q^2_{cv} value of 0.53 (Table 3 model 2a). However, the exclusion of three compounds (**13**, **29**, and **32**) increases model quality to $q^2_{cv} = 0.69$ (s-CoMSA model—Table 3 model 2b). As the experience in 3D QSAR modeling indicates that q^2_{cv} alone cannot be a sufficient indicator of the model quality, we performed further model validation. The asarone series are split into two subseries. The compounds were sampled into these subseries by performing the Kennard–Stone calculations.[45] Then a respective model calculated for first subseries was used for the activity predictions in the second, test group and the SDEP error is calculated. The results and samplings are reported in Table 3. This indicates reasonable predictivity. Figure 2a–d illustrates the relationships between the cross-validated (predicted) and observed values of atherogenic index for the various models reported in Table 3.

Figure 3 (see also Supplementary Materials Figure 2) illustrates the surface areas of the key importance for the compound activity as indicated by s-CoMSA model 2a. We used that model because it includes all compounds. The respective color-coding allows us to identify the influence of the respective points sampled on the molecular surface by the combination of the electrostatic potential value and a value of the b weight in the PLS model. The variables contributing to the activity on a level close to 0 (near 90% of the points sampled) are omitted. Such an illustration suggests the key pharmacophore for the asarones investigated. Thus, an area near the central aromatic ring substituted with an alkoxyl

Table 3. The results of CoMFA and CoMSA 3D QSAR modeling

Compound	Model	q^2_{cv} [a]	r^2 [j]	s	RMS	SDEP
1	CoMFA[b]	0.58	0.71	0.57	—[f]	—[f]
2a	s-CoMSA	0.53	0.76	0.62	0.42	—[f]
2b	s-CoMSA[c]	0.69	0.90	0.45	0.24	—[f]
3a	s-CoMSA-IVE-PLS[g]	0.92	0.98	0.28	0.13	—[f]
3b	s-CoMSA-IVE-PLS[c,h]	0.94	0.97	0.20	0.13	—[f]
4	s-CoMSA	0.55	0.92	0.54	0.19	0.80[d]
5	SOM-CoMSA	0.60	0.90	0.51	0.21	0.64[e]
6a	s-CoMSA[i]	0.45	0.85	0.71	0.34	0.49
6b	s-CoMSA-IVE-PLS[i]	0.75	0.86	0.48	0.33	0.30

[a] LOO cross-validated values; all compounds are included in the model.
[b] Calculation was performed with H(+1) as a probe atom.
[c] After the exclusion of three compounds: **13**, **29**, and **32**.
[d] Test set (K–S subset selection): **1**:**5**, **7**, **11**, **15**, **23**, **28**, **29**, **32**, **36**.
[e] Test set (K–S subset selection): **14**:**17**, **19**, **22**, **24**, **25**, **29**, **32**, **35**, **37**, **38**.
[f] Not tested.
[g] After IVE-PLS starting from model 2a.
[h] After IVE-PLS starting from model 2b.
[i] Test set **33**:**40**.
[j] Fitted statistics.

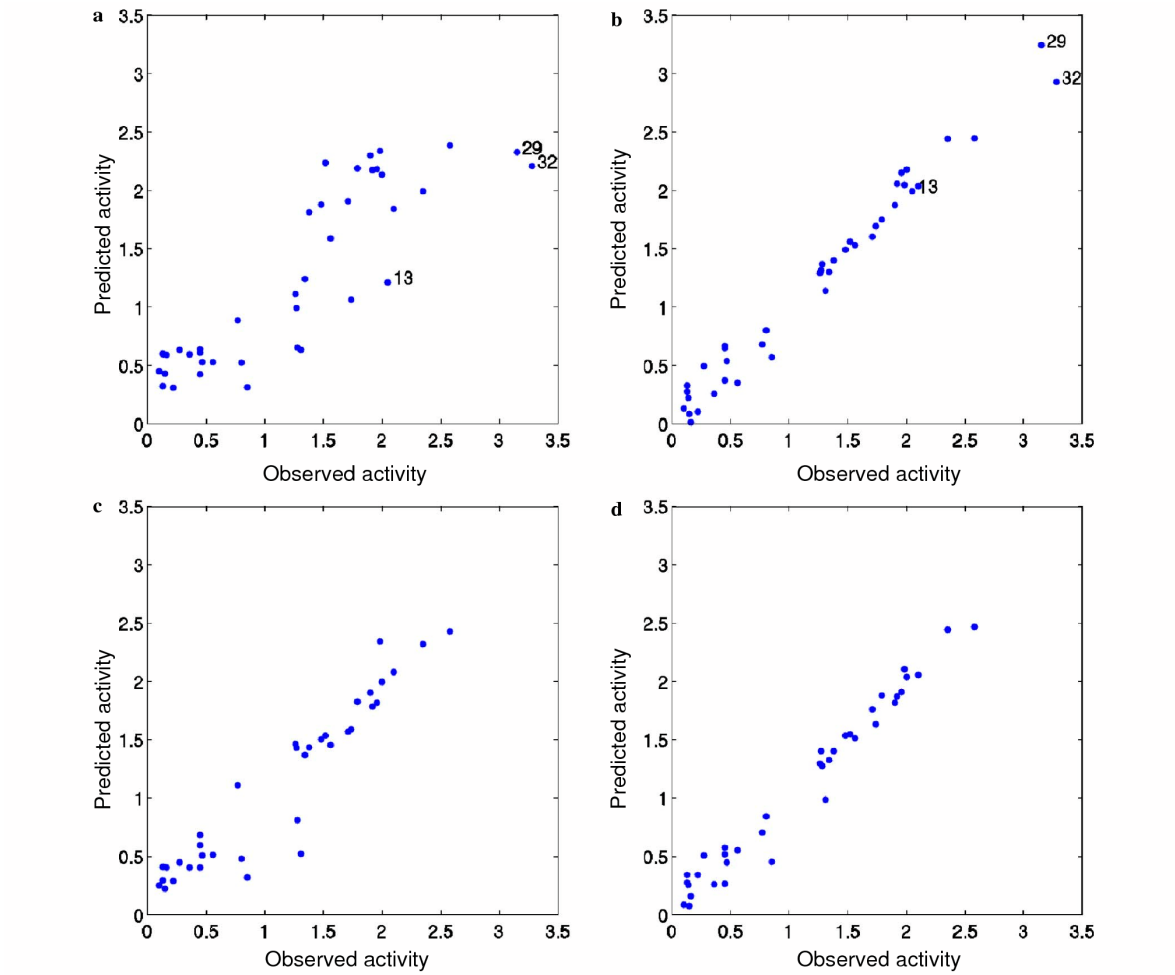


Figure 2. The relationship between the cross-validated (predicted) and observed values of atherogenic index for the various models reported in Table 3; (a) model 2a, (b) model 3a, (c) model 2b and (d) model 3b.

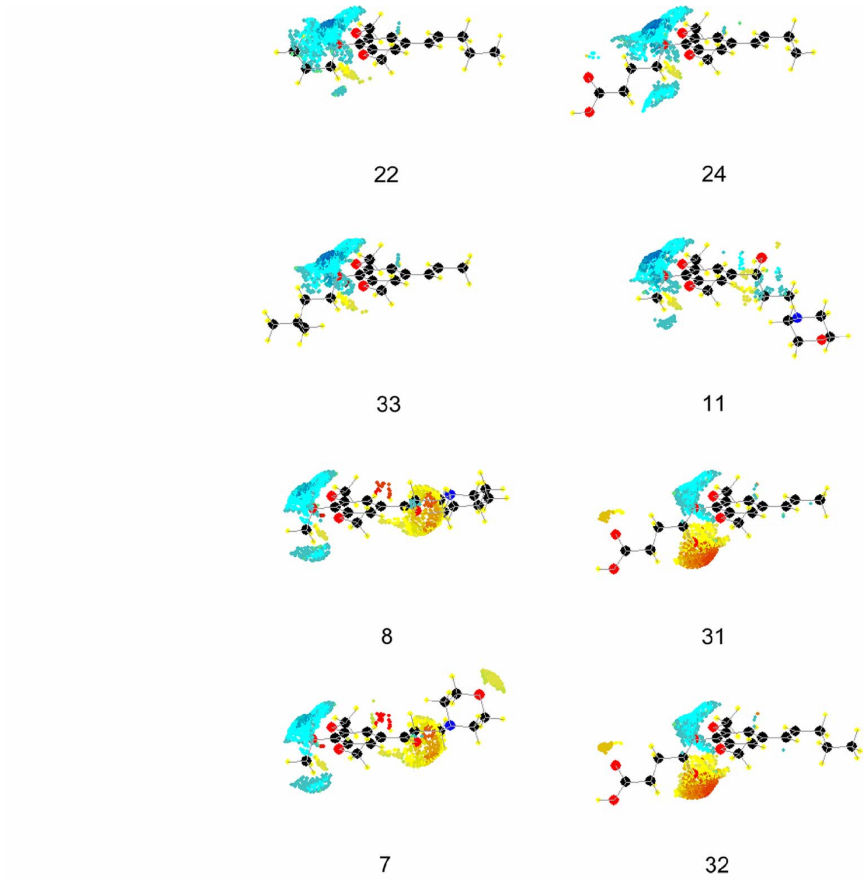


Figure 3. Molecular surface areas of the highest electrostatic contribution to the compound activity according to model 2b, sorted by the decreasing activity of some illustrative molecules (see also Supplementary Materials Figure 2). Blue indicates positive, and yellow and red negative influence on the activity. Details in text.

functionality provides a negative contribution, i.e., increases the activity (blue-colored sections), while a negatively charged carbonyl oxygen in the side chain generally decides a lower activity, clearly decreasing the activity as illustrated by the yellow and red molecular surface areas. This rule can be proved by the examination of the structures given in Table 2. It is worth noticing that if a hydroxyl function replaces a carbonyl oxygen, for example, compound **11**, we do not observe any activity decrease similar to that effected by a carbonyl-group, for example, compound **10**. Thus, our analysis reveals the molecular basis for the hypolipidemic activity of asarones. In Figure 4, we show the results obtained after application of additional filters. This allowed us to obtain highly predictive 3D QSARs (model 3a $q^2_{cv} = 0.92$, $s = 0.28$, RMS = 0.13 and 3b $q^2_{cv} = 0.94$, $s = 0.20$, RMS = 0.13, respectively); however it does not give so clear molecular illustration. Thus, in Figure 4 we give examples of such plots for compound, of high, medium, and low activity. In Figure 4a, the sectors contributing to the activity are divided into two subsets. Green sector increases the activity; orange—decreases. The plots are now more specific for individual compounds, for example, the contribution of the alliloxyl side chain of compound **22** is pronounced more clearly than the areas closer to the aromatic unit. However, the key pharmacophore near the central aromatic unit and the disadvantageous influence of carbonyl oxygen are preserved. Moreover, in Figure 4b we coded by colors the different influences of the electro-negative and -positive surface sectors. This reveals the disadvantageous contribution of the electronegative carbonyl oxygen (dark blue) and electropositive metoxyl (red). The plots shown in Figure 4c allow for the estimation of the relative contribution of the sectors indicated. It is worth noticing that the contour interaction plots displayed in Figures 3 and 4 are completely different from that which resulted from CoMFA modeling. Unlike in CoMFA, each plot is characteristic for individual compound. This enables a clear differentiation of the low and high activity attributes. In particular, this provides also a novel insight into the molecular basis for the asarone activity.

Since overfitting is an important problem in QSAR modeling, we further validated the quality of the CoMSA models by the so-called model randomization. We investigated this by generating 100 random permutations of the activity column. We used model **6a** as an

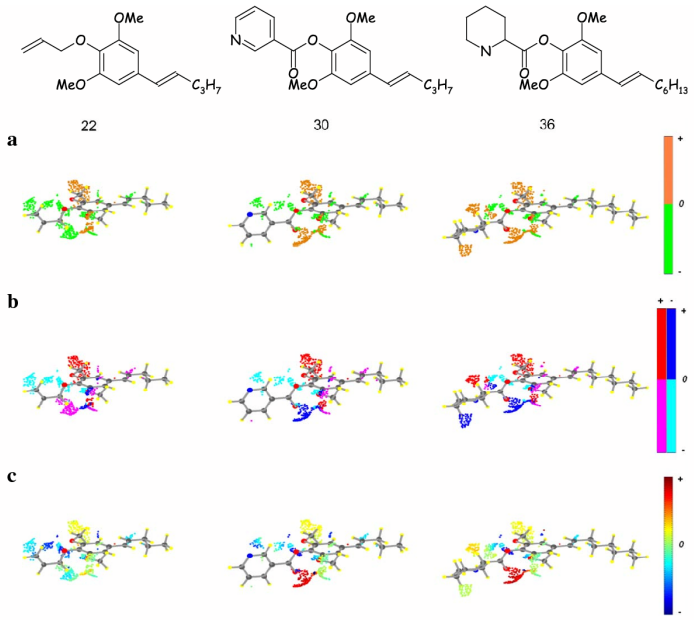


Figure 4. Molecular surface sectors indicated by s-CoMSA-IVE-PLS (model 3b from Table 2) for compounds of the highest, medium, and lowest activity; (a) color codes indicate a sign of activity) change: orange—decreases the activity, green—increases the activity, (b) combination of the electrostatic potential sign and the sign of the *b* weight in the model: +/+ (red, decreases the activity), −/+ (cyan—increases the activity), +/− (magenta, increases the activity), −/− (dark blue, decreases the activity), and (c) color codes indicate a relative value of activity change: warm color (red)—decreases the activity, cool colors (blue)—increases the activity, respectively.

initial form to perform IVE in such random models. Thus, each model with random activity column was processed by the IVE procedure and q^2_{cv} was measured. Figure 5 illustrates the histogram of calculated q^2_{cv}. Red asterisk points to the q^2_{cv} value of the unmodified model, that is, model 6b. Clearly, the majority of models have a relatively low q^2_{cv}, significantly lower than the value calculated for a model with actual compound activity. This validates the statistical significance of the model **6b** where the activity is properly arranged according to the actual activity reported in Table 2.

5. Conclusions

Compounds **7**–**13** exhibited diversified hypolipidemic activity in rats. The most active compound **8** elevated HDL-C level by 35% and diminished TC, LDL-C, and TG levels by 10%, 15.6%, and 16%, respectively (as compared to clofibrate with +24%, −23%, +7%, and −27% respective levels). Although some pseudo- or mini-receptor models were reported for asarones previously, this did not reveal any clear molecular basis for controlling the activity of these compounds. The present analysis employing the CoMSA allowed us to reveal a correlation between the activity of these compounds and the electrostatic potential at the molecular surface. The grid formalism (s-CoMSA) allowed us also to indicate a pharmacophore that is of key importance for the compounds' activity. The CoMSA formalism coupled with the IVE (CoMSA–IVE) allowed us also to obtain highly predictive models.

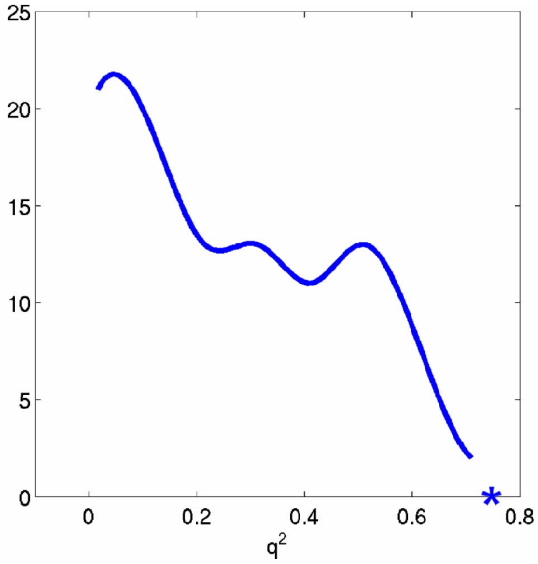


Figure 5. The histogram of q^2_{cv} computed for 100 randomized CoMSA models (random activity arrangement). An asterisk indicates a q^2_{cv} value for the proper activity arrangement. Details in text.

6. Experimental

6.1. Chemical methods and reagents

Melting points were determined with a Köffler apparatus of the Böetius type and are uncorrected. ^{1}H and

^{13}C NMR spectra were recorded on a Bruker AC 200F spectrometer using $CDCl_3$ solution with TMS as internal standard (chemical shifts in δ ppm). IR spectra were recorded on a Nicolet Magna 550 FTIR Spectrometer in chloroform solutions. The UV spectra were collected on a Hewlett-Packard UV–vis Diode Array Spectrophotometer 8452A. Mass spectra were obtained at 70 eV with an AMD-604 spectrometer. The reaction products were isolated by column chromatography or flash chromatography performed on a silica gel 70–230 mesh ASTM (Merck, Darmstad, Germany). Thin-layer chromatograms were developed on aluminum TLC sheets precoated with silica gel F_{254} (Merck, Darmstadt, Germany). The spots were visualized with 50% sulfuric acid after heating. All the solvents were dried and freshly distilled prior to use. Starting materials and reagents were purchased from Aldrich Chemical Co. (Steinheim, Germany).

6.2. General procedure for 3-amin-4-yl-1-(3,4,5-trimethoxyphenyl)propan-1-one (7–10)

The mixture of 0.046 mol of hydrochloride cycloamine, 2.0 g (0.066 mol) of paraformaldehyde, 2 g 3,4,5-trimethoxyacetophenone (0.040 mol), 2.4 mL concd HCl, and 20 mL absolute ethanol in a 100 mL three-necked flask was refluxed for 2 h. Then 1.2 g (0.040 mol) paraformaldehyde was added. The reaction mixture was refluxed for 5 h. To a hot mixture 50 ml acetone was added and refluxed for 15 min. After cooling down to room temperature, the reaction mixture was evaporated under reduced pressure to dryness. The product was crystallized from ethanol/acetone 9:1 (v/v).

6.3. 3-Morpholin-4-yl-1-(3,4,5-trimethoxyphenyl)propan-1-one (7)

Mp 199–201 °C; IR ($CHCl_3$, cm^{-1}) *v*: 2700–2400; 1697; 1200; 1118; 1131; ^{1}H NMR ($CDCl_3$, 200 MHz) δ: 11.3 (s, 1H); 7.2 (s, 2H); 4.2 (m, 2H); 4.1 (m, 2H); 3.9 (s, 9H); 3.7 (m, 2H); 3.4 (m, 4H); 2.9 (m, 2H); ^{13}C NMR ($CDCl_3$, 50 MHz) δ: 194 (C); 153 (C); 143 (C); 130 (C); 105 (CH); 63 (CH_2); 60 (CH_3); 56 (CH_3); 52 (CH_2); 32 (CH_2); MS *m/z* 100 (M^+, 100).

6.4. 3-(4-Methylpiperidin-1-yl)-1-(3,4,5-trimethoxyphenyl)propan-1-one (8)

Mp 153–155 °C; IR ($CHCl_3$, cm^{-1}) *v*: 2700–2400; 1684, 1200, 1133; ^{1}H NMR ($CDCl_3$, 200 MHz) δ: 12.2 (s, 1H); 7.3 (s, 2H); 3.9 (s, 6H); 3.8 (s, 3H); 3.7 (t, 2H); 3.5 (m, 4H); 2.7 (m, 2H); 1.8 (m, 5H); 1.1 (d, 2H); ^{13}C NMR ($CDCl_3$, 50 MHz) δ: 194 (C); 152 (C); 142 (C); 130 (C); 102 (CH); 66 (CH_2); 60 (CH_3); 56 (CH_3); 52 (CH_2); 32 (CH_2); MS *m/z* 112 (M^+, 100).

6.5. 3-Azepan-1-yl-1-(3,4,5-trimethoxyphenyl)propan-1-one (9)

Mp 171–173 °C; IR ($CHCl_3$, cm^{-1}) *v*: 2700–2400; 1680, 1202, 1131, 1131; ^{1}H NMR ($CDCl_3$, 200 MHz) δ: 12.2 (s, 1H); 7.2 (s, 2H); 3.9 (s, 6H); 3.8 (s, 3H); 3.7 (m, 2H); 3.6 (m, 4H); 2.9 (m, 2H); 1.8 (m, 4H); 1.6 (m, 2H); ^{13}C NMR ($CDCl_3$, 50 MHz) δ: 194 (C), 152 (C); 142 (C); 130 (C); 105 (CH); 52 (CH_2); 60 (CH_3); 56 (CH_3); 45 (CH_2); 33 (CH_2); 26 (CH_2); 24 (CH_2); 22 (CH_2); MS *m/z* 112 (M^+, 100).

6.6. 3-Piperidin-1-yl-1-(3,4,5-trimethoxyphenyl)propan-1-one (10)

Mp 203–204 °C; IR ($CHCl_3$) *v* 2700–2400; 1697, 1200, 1131 cm^{-1}, ^{1}H NMR ($CDCl_3$, 200 MHz) δ: 11.4 (s, 1H); 7.2 (s, 2H); 3.9 (s, 6H); 3.8 (s, 3H); 3.7 (m, 2H) 3.6 (m, 4H); 2.6 (m, 2H); 2.1 (m, 2H); 2.1 (m, 2H); 1.7 (t, 2H); 1.5 (m, 2H); ^{13}C NMR ($CDCl_3$, 50 MHz) δ: 194 (C); 152 (C); 142 (C); 130 (C); 105 (CH); 53 (CH_2); 60 (CH_3); 56 (CH_3); 32 (CH_2); 22 (CH_2); 21 (CH_2); MS *m/z* 112 (M^+, 100).

6.7. Synthesis of 3-morpholin-4-yl-1-(3,4,5-trimethoxyphenyl)propan-1-ol (11) 3-piperidin-1-yl-1-(3,4,5-trimethoxyphenyl)propan-1-ol (12)

3.5 g (0.010 mol) of hydrochloride 3-morpholin-4-yl-1-(3,4,5-trimethoxyphenyl)propan-1-one or 3-piperidin-1-yl-1-(3,4,5-trimethoxyphenyl)propan-1-one was dissolved in 2 ml water. After adding (20 mL) 6 M NaOH, the mixture was extracted three times with ether (3× 30 mL). The ether extract was dried over $MgSO_4$ and evaporated in vacuo. The crude product was dissolved in 30 mL methanol. The reaction mixture was dropped to a cold mixture of 440 mg (0.010 mol) $NaBH_4$ dissolved in 20 mL methanol and 20 mL water. The reaction mixture was stirred for 2 h at room temperature and 15 min at 45 °C. The organic solvent was removed and 15 ml of 6 N NaOH was added. The aqueous solution was extracted with chloroform (3× 30 mL). The organic extract were combined and washed with water and dried over $MgSO_4$. The mixture was filtered off and concentrated under a reduced pressure. After evaporation of solvent, the residue was purified by silica gel column chromatography (elution with hexane/ether; 1:1; v/v) and crystallized from an ethanol/acetone 9:1 (v/v) .W = 63.5%.

6.8. 3-Piperidin-1-yl-1-(3,4,5-trimethoxyphenyl)propan-1-ol

Mp 65–68 °C; IR ($CHCl_3$, cm^{-1}) *v*: 3181, 1227, 1132, 1112; ^{1}H NMR ($CDCl_3$, 200 Hz) δ: 6.6 (s, 2H); 4.9 (t, 1H); 3.9 (s, 6H); 3.8 (s, 6H); 2.7 (m, 4H); 2,4 (m, 2H); 1.9 (m, 2H); 1.6 (m, 4H); 1.5 (m, 2H); ^{13}C NMR ($CDCl_3$, 50 Hz) δ: 153 (C); 139 (C); 136 (C); 106 (CH); 80 (CH); 54 (CH_2); 52 (CH_2); 33 (CH_2); 22 (CH_2); 60 (CH_3); 57 (CH_3); MS *m/z* 98 (M^+, 100).

6.9. 3-Morpholin-4-yl-1-(3,4,5-trimethoxyphenyl)propan-1-ol

Mp 86–88 °C; IR ($CHCl_3$) *v*: 3190, 1227, 1132, 1119, 1127 cm^{-1}; ^{1}H NMR ($CDCl_3$, 200 MHz) δ: 6.6 (s, 2H); 4.9 (t, 1H); 3.9 (s, 9H); 3.7 (t, 4H); 2.7 (m, 4H); 2.5 (m, 4H); 1.9 (m, 2H); ^{13}C NMR ($CDCl_3$, 50 MHz) δ: 154 (C); 140 (C); 138 (C); 102 (CH); 75 (CH); 68 (CH_2); 61 (CH_3); 58 (CH_2); 57 (CH_3); 53 (CH_2); 33(CH_2); MS *m/z* 100 (M^+, 100).

6.10. Synthesis of 4-[(2*E*)-3-(3,4,5-trimethoxyphenyl)prop-2-enyl]morpholine (13)

A solution of crude alcohol, 2.7 g (0.009 mol) 3-morpholin-4-yl-1-(3,4,5-trimethoxyphenyl)propan-1-ol in 50 ml of dry toluene was treated with 3 g 0.02 mol of anhyd $CuSO_4$. The reaction mixture was heated for 2 h at reflux. After the evaporation of the solvent, the residue was crystallized from ethanol/acetone 9:1 (v/v).

Mp 85–86 °C; IR ($CHCl_3$, cm^{-1}) *v*: 3181, 1227, 1132, 1112; ^1H NMR ($CDCl_3$, 200 Hz) *δ*: 6.86 (s, 2H); 6.1 (d, 1H, $J = 15.8$ Hz); 5.9 (dt, 1H, $J = 6.6$ Hz, $J = 15.75$ Hz); 3,85 (s, 6H); 3,8 (s, 6H); 3,38 (m, 4H); 2,2 (dt, 2H, $J = 6.65$ Hz, $J = 7{,}04$ Hz); 2.1 (m, 2H); ^{13}C NMR ($CDCl_3$, 50 MHz) *δ*: 154 (C); 138 (C); 133 (CH); 128 (CH); 120 (C), 104 (CH); 69 (CH_2); 66 (CH_2); 60 (CH_3); 56 (CH_3); 53 (CH_2).

Acknowledgments

The authors thank Professor Johann Gasteiger of the University of Erlangen-Nürnberg, Germany, both for facilitating access to the CORINA, PETRA, SURFACE, and KMAP programs. The financial support of the KBN Warsaw under Grants No. KBN P05F 01617, 4P05F019 19, KBN 4T09A 088 25, and PBZ 040 P04/08 is gratefully acknowledged. R.G. thanks the Foundation for Polish Science for an individual grant.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2005.10.014.

References and notes

1. Anderson, K. M.; Wilson, P. W. F.; Odell, P. M.; Kannel, W. B. *Circulation* **1991**, *83*, 356.
2. Van Gaal, L. F.; Hang, A.; Steijaert, Mm.; De Leeuw, Ih. *Int. J. Obes.* **1995**, *19*, S21.
3. McCully, K. S. *Nat. Med.* **1996**, *2*, 386.
4. Brown, Bg.; Zhao, X.; Sacco, De.; Albers, J. J. *Circulation* **1993**, *87*, 1781.
5. Badimon, J. J.; Fuster, V.; Badimon, L. *Circulation* **1992**, *86*, II-86–III-94.
6. Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III) *JAMA* **2001**, 285, 2486.
7. Chilmonczyk, Z.; Siluk, D.; Kaliszan, R. *Exp. Opin. Ther. Pat.* **2001**, *11*, 1301.
8. Levine, G.; Keaney, J.; Vita, J. *N. Engl. J. Med.* **1995**, *332*, 512.
9. Szapary, P. O.; Rader, D. J. *Am. Heart. J.* **2004**, *148*, 211.
10. Eghdamian, B.; Ghose, K. *Drugs Today* **1988**, *35*, 79.
11. Farmer, J. A.; Gotto, A., Jr. *Adv. Pharmacol.* **1966**, *35*, 79.
12. Enriquez, R.; Chávez, M.; Jáuregui, F. *Phytochemistry* **1980**, *19*, 2024.
13. Chamorro, G.; Garduno, L.; Sanchez, A.; Labarrios, F.; Salazar, M.; Martinez, E.; Diaz, F.; Tamariz, J. *Drug Dev. Res.* **1998**, *43*, 105.
14. Labarrios, F.; Garduno, L.; Vidal, M.; Garcia, R.; Salazar, M.; Martinez, E.; Diaz, F.; Chamorro, G.; Tamariz, J. *J. Pharm. Pharmacol.* **1999**, *51*, 1.
15. Dandiya, P. C.; Sharma, J. D. *Ind. J. Med. Res.* **1962**, *50*, 46.
16. Dandiya, P. C.; Menon, M. K. *Br. J. Pharmacol.* **1963**, *20*, 436.
17. Belova, L.; Alibekov, S.; Baginskaya, A.; Sokolov, S.; Pokrovskaya, G.; Stikhin, V.; Trumpe, T.; Gorodnyuk, T. *Farmak. Toksikol.* **1985**, *48*, 17.
18. Morales, R.; Madrigal, B.; Mercader, M.; Cassani, M.; Gonzalez, G.; Chamorro, G.; Salazar, M. *Mutat. Res.* **1992**, *279*, 269.
19. Salazar, M.; Salazar, S.; Ulloa, V.; Mendoza, T.; Pages, N.; Chamorro, G. *J. Toxicol. Clin. Exp.* **1992**, *12*, 149.
20. Diaz, F.; Contreras, L.; Flores, R.; Tamariz, J.; Labarrios, F.; Chamorro, G.; Munoz, H. *Org. Prep. Proc. Int.* **1992**, *24*, 78.
21. Filipek, S.; Łozowicka, B. *Acta Pol. Pharm.* **2000**, *57*, 106.
22. Cruz, M.; Salazar, M.; Garciafigueroa, Y.; Hernandez, D.; Diaz, F.; Chamorro, G.; Tamariz, J. *Drug Dev. Res.* **2003**, *60*, 186.
23. Cruz, A.; Garduno, L.; Salazar, M.; Martinez, E.; Jimenez-Vazquez, H.; Diaz, F.; Chamorro, G.; Tamariz, J. *Arzneim.-Forsch./Drug Res.* **2001**, *51*, 535.
24. Poplawski, J.; Lozowicka, B.; Dubis, A.; Lachowska, B.; Witkowski, S.; Siluk, D.; Petrusewicz, J.; Kaliszan, R.; Cybulski, J.; Chilmonczyk, Z.; Strzalkowska, M. *J. Med. Chem.* **2000**, *43*, 3671.
25. Lozowicka, B.; Filipek, S.; Dubis, A. *Eur. J. Med. Chem.* (submitted).
26. Lozowicka, B.; Poplawski, J.; Dubis, A.; Chilmonczyk, Z.; Cybulski, J.; Kita, K.; Kobes, S.; Filipek, S. *Drug Dev. Res.* (submitted).
27. Chilmonczyk, Z.; Siluk, D.; Kaliszan, R.; Lozowicka, B.; Poplawski, J.; Filipek, S. *Pure Appl. Chem.* **2001**, *73*, 1445.
28. Polanski, J.; Walczak, B. *Comp. Chem.* **2000**, *24*, 615.
29. Polanski, J.; Gieleciak, R.; Bak, A. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 184.
30. Polanski, J.; Gieleciak, R. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 656.
31. Polanski, J.; Gieleciak, R.; Wyszomirski, M. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1754.
32. Polanski, J.; Gieleciak, R.; Wyszomirski, M. *Dyes Pigments* **2004**, *62*, 63.
33. Polanski, J.; Gieleciak, R. *Mol. Diversity* **2003**, *7*, 45.
34. Polanski, J.; Gieleciak, R.; Bak, A. *Comb. Chem. High Throughput Screen.* **2004**, *7*, 793.
35. Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. *Molecules* **2004**, *9*, 1148.
36. Polanski, J.; Gieleciak, R.; Magdziarz, T.; Bak, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1423.
37. MATLAB 5.0 program. Available from: The Mathworks Inc., Natick, MA, USA, http://www.mathworks.com.
38. Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandeginste, B. M. V.; Sterna, C. *Anal. Chim. Acta* **1996**, *330*, 1.
39. Polanski, J.; Bak, A. J. *Chem. Inf. Comput. Sci.* **2003**, *43*, 208.
40. Blicke, F.; Bruckhalter, J. *J. Am. Chem. Soc.* **1942**, *64*, 451.
41. Testa, B.; Purcell, W. P. *Eur. J. Med. Chem.* **1978**, *13*, 509; Motoc, I. Molecular Shape Descriptors. In *Steric Effects in Drug Design*; Charton, M., Motoc, I., Eds.; Akademie: Berlin, 1983; pp 93–105.
42. Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatsu, K. *Comput. Chem.* **2002**, *26*, 583.

43. Hasegawa, K.; Morikami, K.; Shiratori, Y.; Ohtsuka, T.; Aoki, Y.; Shimma, N. *Chemometr. Intell. Lab. Syst.* **2003**, *69*, 51.
44. Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatsu, K. *Comput. Biol. Chem.* **2003**, *27*, 381.
45. Kennard, R. W.; Stone, L. A. *Technometrics* **1969**, *11*, 137.

Modeling Robust QSAR†

Jaroslaw Polanski,* Andrzej Bak, Rafal Gieleciak, and Tomasz Magdziarz

Department of Organic Chemistry, Institute of Chemistry, University of Silesia, PL-40-006 Katowice, Poland

Received August 5, 2005

Quantitative Structure Activity Relationship (QSAR) is a term describing a variety of approaches that are of substantial interest for chemistry. This method can be defined as indirect molecular design by the iterative sampling of the chemical compounds space to optimize a certain property and thus indirectly design the molecular structure having this property. However, modeling the interactions of chemical molecules in biological systems provides highly noisy data, which make predictions a roulette risk. In this paper we briefly review the origins for this noise, particularly in multidimensional QSAR. This was classified as the data, superimposition, molecular similarity, conformational, and molecular recognition noise. We also indicated possible robust answers that can improve modeling and predictive ability of QSAR, especially the self-organizing mapping of molecular objects, in particular, the molecular surfaces, a method that was brought into chemistry by Gasteiger and Zupan.

INTRODUCTION

It may look like a paradox but *the most fundamental and lasting objective of* (chemical) *synthesis is not a production of new compounds but the production of properties.*[1] However, a problem is that property production is still more of a dream than a reality. Despite the fact that the recent decade has brought forth a number of revolutionary ideas in molecular design, e.g. combinatorial chemistry, genomics, chemogenomics, the number of newly registered drugs decreases.[2] What is the place of QSAR in a novel molecular design landscape?

Basically, QSAR should work like a dictionary between the chemical compound space and the property space. This method includes a variety of procedures that usually starts from searching mathematical models describing a biological answer in a given property space. Next, the compound space is screened in a search for the virtual molecules of the required property. Virtual objects having promising properties can be synthesized in a hope for the optimization of the molecular structure. Therefore, we can define QSAR as an indirect molecular design by the iterative sampling of the chemical compound space to optimize a certain property and thus indirectly design the molecular structure having this property. Although generally QSAR realizes a strategy from molecules to property, the so-called inverse strategy from property to molecules has also been investigated.[3]

QSAR Is Highly Data Dependent. QSAR relates chemical or biological answers to the molecular structure offering a very different level of the explanation of the mechanisms controlling real processes. The Hammett equation was a first quantitative structure−reactivity approach in chemistry. Generally, it fails however to model biological activity. Hansch developed the first successful QSAR model that describes biological responses by including a hydrophobicity (log P) term.[4] A linear equation which takes the form of $\log(1/C) = a\log(P) + C$ is the simplest function that can relate structure, or more accurately a calculable property manifested by such a structure, to activity connecting the activity, log(1/C), and property, log(P) spaces. In fact, a number of QSAR models described by a similar equation can be found in the literature. However, generally log(P) is a parameter that describes biological transport phenomena. From the theoretical point of view this implies that there is some optimal log(P) value, and a linear function cannot be used for modeling such a phenomenon. Figure 1 illustrates a data dependency problem in a situation when linear and parabolic models can describe hypothetical activity. The effect presented in Figure 1 brings a problem of the predictive credibility. Neither the linear model **a** nor the linear model **b** is capable of the proper Y* predictions in space X, for the nonlinear Y vs X relationship. Figure 1a,b shows also a fact that the extrapolation of linear models is the most dangerous for the prediction credibility. We should however keep in mind that the biological activity profile in the molecular structure landscape is *controlled by the similarity paradox,* i.e., even a minute change in the compound structure can result in a substantial activity change. This makes any QSAR prediction for **a real external object** (a virtual molecule that has not been synthesized before the modeling step) an extrapolation beyond the well-explored borders rather than an interpolation. In this context, making predictions, which seems to be a main objective of QSAR, is a roulette risk operation.

Robust Answers for QSAR Data Dependence. QSAR is not the only example of data dependency in modeling. Weather forecasts can be an illustrative example of the highly risky predictions under extremely unstable conditions originated from a large number of hard to control variables. The reduction of the sensitivity of the modeled output to the variation of inputs that should decrease data dependency and improve predictions in such conditions is called *robust modeling. The term robustness in signal processing applications usually refers to approaches that are not degraded*

† Dedicated to Professor Johann Gasteiger.

* Corresponding author e-mail: polanski@us.edu.pl.

10.1021/ci050314b CCC: $33.50
Published on Web 12/08/2005

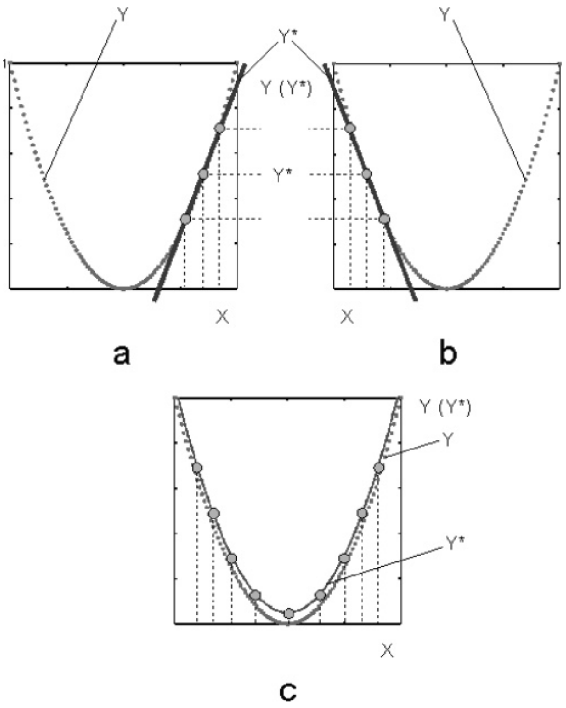


Figure 1. A schematic view of data dependency. In this particular case a difference between the modeled (Y*) and real (Y) answers are controlled by the discrepancy between a mathematical form of the equation used. If we assume that Y is given by a parabola in the X space, then a linear equation can be approximated by the assumed parabolic Y answer only in some limited X ranges (a) or (b). However, we need to include a nonlinear term to describe a parabola (c). Models (a) or (b) will be highly data dependent, and their application for the predictions of Y* will be restricted only to the certain X data range.

significantly when the assumptions that were invoked in defining the processing algorithm are no longer valid.[5]

Data dependency is well realized in QSAR, and a controversy about reliability of similar approaches in the context of predictability and practical applications among the medicinal chemistry audience is alive and well. The most robust answer would be almost completely data independent. Recently, several approaches have appeared in drug design that can be included in such a trend. Thus, the Lipinski rule of five,[6] druglikeness,[7] comparative QSAR, combined QSAR databases,[8,9] or the ADMET approach[10] are the methods that are based on the search for common features that connect all drugs. Oprea tested more than 12 000 compounds of the different biological activity types attempting to find the clusters of active, middle-activity, and inactive molecules. In a similar approach the relationship of the regression coefficients for more than 400 QSAR have been investigated.[11] Formally, in similar approaches we search for a certain property range, rather than a discrete property value, that would describe a hypothetical virtual drug space.

The *extensive data independence* implies, however, qualitative and not quantitative solutions. What are the solutions for quantitative modeling? The replacement of the linear Hansch QSAR model by the incorporation of the squared (log P)[2,4] or bilinear[12] terms indicates possible directions of the improvement of the model robustness, in this particular case, by eliminating the discrepancy between real interaction mechanism and mathematical formula we can model a single equation relating broader X range to Y. New computational methods including neural networks, data elimination, genetic algorithms, and novel model validation schemes are other examples in this field.[8,13−15] A combination of different data handling schemes as neural networks and genetic algorithms or neural networks and PLS analysis appeared to be especially effective. PLS analysis is a default method used in a number of 3D QSAR methods.[16] Moreover, modeling equations can be supported by more flexible data handling methods. Thus, for example, Support Vector Machine (SVM) is a new promising method for data classification and regression that has recently gained special attention in many fields of chemistry and medicine.[17,18] It has been introduced as a robust and highly accurate intelligent classification technique, well suited for QSAR applications.[19] Recent SVM modifications i.e., Support Vector Regression (SVR), offer a good prediction performance and can be used as the PLS replacement.[20] An interesting comparison between this technique and other QSAR tools can be found in ref 21. This method appeared especially useful in QSAR when data are not linearly separable. In this particular case, SVM can classify properly in a linear way by constructing the Optimal Separating Hyperplane (OSH) in a space of higher dimensionality. To avoid time-consuming calculations space mapping is realized implicit by so-called kernel functions.[22] Instead of minimizing the error on the training data the SVM maximizes the margin, i.e, the largest possible distance from the hyperplane to the closest objects of the two classes. The solution to the optimization problem is a global minimum, whereas other machine learning methods sometimes terminate in a local minima. In comparison to other QSAR techniques SVM gives the similar results of analysis but requires significantly smaller training times, which is increasingly important when learning large numbers of chemical compounds.[23]

Robustness and Data Noise. Usually, in QSAR data describing no more than 100 compounds are available. In traditional QSAR the activity of these molecules are related to several molecular descriptors. This situation dramatically changes in multidimensional QSAR where the number of molecular descriptors increases to thousands. Standard regression fails to deal with such data structure due to a number of cross-dependencies and chance correlations.[24] A robust answer for the data flood condition, i.e., data noise, can be Principal Component Analysis (PCA) or Partial Least Squares (PLS) methods that attempt to get better insight into the molecular descriptor space by data projection to so-called latent variables which are linear combinations of original variables. This problem will not be discussed here further, and a reader is referred to a number of monographs available.[25−27] Data noise demands special model validation schemes since fitted statistics applied for standard regression fail to provide a reliable analysis of the model quality. Cross-validation (CV) that includes several techniques such as k-fold leave-one-out (LOO), jackknife, delete-d, or bootstrapping[28] can be applied for the PLS model validation. Essentially, as shown in Figure 2, CV describes a procedure in which we divide the data into a number of groups to develop a number of parallel models, while one of the groups is deleted.[29−31] The activity predicted for this group is then

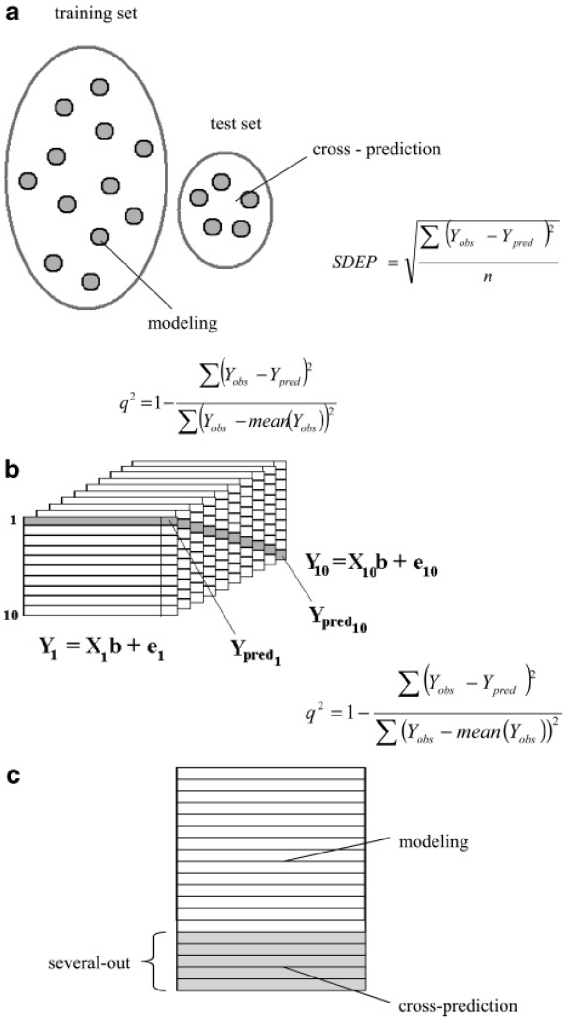


Figure 2. A validation of data noisy multidimensional QSAR models by cross-prediction. The sampling within molecular objects forms the *training* and *test,* respectively (a). Usually, an iterative leave-one-out cross-validation LOO CV (b) or cross-prediction by a single LSO/LOO CV iteration (c) is used for the estimation of the model relevance. Thus, in a series of experiments a model calculated for *all but one* molecule is used for the calculation of the activity of the eliminated molecule (b) or a model calculated for the *training* set (e.g., molecules 1−7) is to be cross-predicted into the *test* set (e.g., molecules 8−10) (c). The q^2 or SDEP (r^2) parameters are used for measuring model quality, respectively. $q^2 = 1 - (\sum(\text{obs}_i - \text{pred}_i)^2/\sum(\text{obs}_i - \text{mean(obs)})^2)$ SDEP $= (\sum(\text{pred}_i - \text{obs}_i)^2/n)$ where obs is the assayed values; pred is the predicted values, mean is the mean value of obs, and i refers to the object index, which ranges from 1 to m, n − the number of compounds included in the test series.

used for validating final model quality. Often, the deleted group, or so-called test set, contains a single molecule. Iterative manyfold cross-prediction for all molecules using the models derived from *all with the exception of this one molecule* is called leave-one-out (LOO) CV (Figure 2b). In practice, a q^2 parameter calculated during LOO CV is a basic measure for the performance of multidimensional QSAR models. Alternatively, if a larger test set is defined among molecular objects, the model quality is tested by activity predictions within this set by a model optimized using all but test set molecules, i.e, so-called training set molecules. As a rule, a single test set combination is tested in the latter process, and an SDEP parameter probes model quality, as illustrated in Figure 2c. Since, in this particular case, modeling within the training set must also be performed using the robust LOO CV protocol, the latter procedure can be formally described as a single iteration of the leave-some (or several)-out (LSO) CV coupled with LOO CV (LSO/LOO). Tropsha observed that, if we test several LSO/LOO iterations, no correlation between the quality of the model (given by a q^2 value) and the quality of prediction in the test set (SDEP values) exists.[32−34] In other words, the high quality model generated in the training set does not guarantee a proper prediction in the test set. The reasons for that can be easily explained.[35] Thus, during model validation in 3D-QSAR we are processing a strictly limited set of molecules for which an activity has been measured a priori. All of them should be active compounds. In fact, none of the predictions are real external predictions toward virtual molecules of really unknown properties. Thus, we propose in this publication to use for such a condition a term cross-prediction. Intuitively, during cross-prediction we can indicate at least two types of molecular objects: these of higher congenericity (hypothetical similarity level implying similar ligand−receptor interaction mechanism) that can fit easily a common mathematical model, and those of the lower fit into such a model. If we select into the training set preferentially the objects that can be easier fitted into the model, then the chance that the remaining group would provide a worse fit is higher since only the *worse* molecules are available for the test set. Because we have not made any provisions for the division of the molecules into the training and test sets any correlation between modeling ability in the training set and cross-predictions in the test set would be rather a surprise.

Stochastic Model Validation (SMV) for the Data Noise Conditions. To explore further cross-prediction we generalized model validation by the investigations into all possible combinations of the molecular objects into the training/test containing t (training) and *all-t* (test) set molecular objects, respectively. This will give *all!/(all-t)!t!* such combinations. Technically, the scheme is constructed by probing all LSO/LOO iterations, as shown in Figure 3. Formally, this method is an extension of the single iterations analyzed by Tropsha or Doweyko.[32−34] The application of SMV for the analysis of the modeling and cross-predictive ability of QSAR schemes has been discussed thoroughly in ref 35. Figure 4 provides a few illustrative examples of this issue. Since the SMV scheme analyzes the influence of all inputs upon the quality of the cross-predictive models, it should be a nice measure of the modeling robustness, and we will use it below for such purposes. The main conclusion from the SMV is *that the q^2 estimator measures modeling rather than predictive ability*. Moreover, *modeling ability in the training set and predictive capability in the test set are of the dichotic nature, i.e., the higher modeling ability of the training group the lower predictability in a test set*.[28]

If we test a given QSAR data by a single training/test set sampling, then we obtain two single values of the SDEP and q^2 estimators characterizing the prediction/modeling ability within these two sets. However, because both these estimators depend on the way in which we sample the data, the

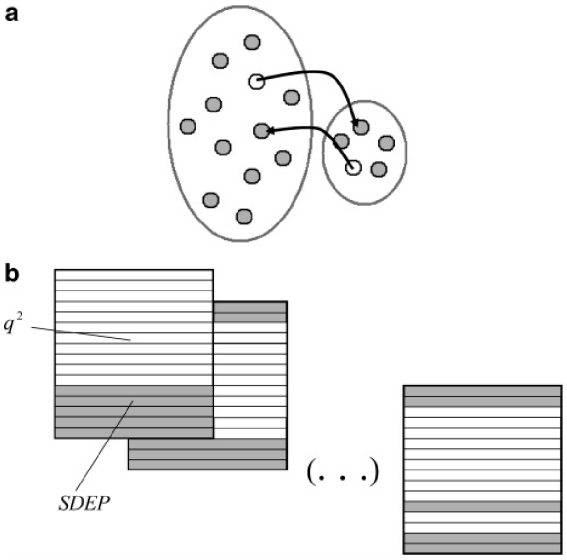


Figure 3. Stochastic model validation. The molecules are iteratively shuffled between the training and test sets until all combinations are probed (a) and a coupled LSO/LOO CV is iteratively repeated (b).

quality of the model can only be evaluated by the comparison of these two estimators. In contarary, SMV can be interpreted as a data probing technique. Thus, by changing the training/test set data sampling we are disturbing the statical QSAR modeling technique and observing how the answer of the model looks like when we change the modeling basis. For a given bioactive compound series the higher the robustness of the model is the lower a diffusion of an ideal single point answer is (Figure 4a).

The analysis of the QSAR data by cross-prediction in stochastic schemes probing intensively larger model populations, i.e., of the protocols similar to SMV, can be found in the literature only very rarely. To estimate the predictive ability more precisely Clark suggested a method he called *boosted leave-many-out CV*.[36] This method estimates model quality by the calculation of the external predictive error of the models obtained by several divisions of the analyzed series into smaller training sets and larger test sets. Training and test sets are selected using the OptiSim procedure which can generate representative or diverse test sets. The analysis of the external and the internal prediction ability of the obtained models confirms its dichotic nature. Sheridan who investigated the dependence of the selection of the training set on the prediction accuracy developed so-called retrospective CV.[37] Thus, a user-specified number of molecules are randomly selected from the original data set or a subset therefrom. These molecules form a training set. Molecules that are not included in the training set form the test set. Then, a QSAR model is generated within the training set, and the model is used to predict the activity of all molecules in the original data and a note is made of which were in the training set. Each molecule in the original data set is assigned an extrapolation measure of how close it is to the training set. Then, the procedure is repeated 10 times. The results were displayed as the predicted vs observed activity plots. The main conclusion is that similarity to molecules in the training set is a good discriminator for the accuracy of the QSAR cross-predictions.

Robust Predictions in Virtual Chemical Compound Space. In fact, cross-predictions in multidimensional QSAR are performed not in the direct search for new molecules, i.e., for molecular design, but for model validation. It is naive nowadays to expect a reliable prediction for a single **virtual molecule** on the basis of multidimensional QSAR. Instead, in this method we use the equations optimized by robust model validation for the illustration of the so-called interaction contour plots. In standard CoMFA the interaction surfaces are determined by filtering regression weights that decide a contribution of original variables (potential values in the grid points) into a final PLS model. The points of the highest standard deviation for the whole molecular series form the space sectors of positive and negative influence for the activity. Basically, an extension of the molecules into such regions should increase or decrease compound activity. Compare refs 38 and 39 for the examples of such a molecular

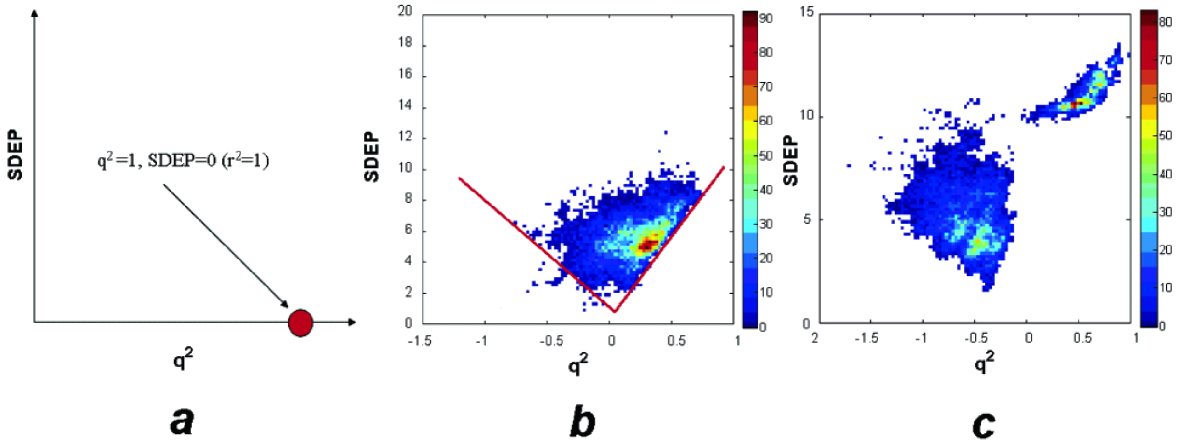


Figure 4. Stochastic model validation for simulated data. If we assume an ideal system in which the Y answer can always be given by modeled Y* without any error (Y*=Y), then a single point reports the SMV experiment ($q^2 = 1$, SDEP = 0) (a). The inclusion of the noise (Y*=Y + noise) results in the explosion of the SMV probes into the q^2, SDEP plot (b). If two different models operate for Y*, i.e., $Y_1^* = f(X)$ or $Y_2^* = g(X)$, then binomial plot forms a response in SMV. Adapted from ref 35.

design (prediction) approaches procedure.

Robust Answers for Molecular Superimposition Noise. Two different points are the only systems that can be put together without any ambiguity. Generally, for two systems that are not identical a number of covering modes exists. As a result superimposition generates an important noise that comes into the modeling. In consequence, in 3D-QSAR, and in CoMFA in particular, molecular superimposition significantly influences final results, e.g., the interaction contour plots can be completely different for two different superimposition modes (see ref 34 for further discussion).

The uncertainty in this context is due to a question if we can cover certain atoms within two molecules, respectively. The most radical solution that completely reduces the uncertainty due to molecular superimposition is passing over this operation.[40] Formally, this can be achieved by using some superimposition invariants as *distance geometry*,[41] *autocorrelation vectors*,[42] *3D MoRSE or RDF codes*,[43] or by the application of some default covering modes, e.g. covering along the molecular inertial axis in the CoMSIA,[44] Receptor-like Neuron Network,[45] GRIND,[46] or CoMMA[47] methods. Although we usually do not realize this, in fact, default superimposition is also performed by a variety of 2D-QSARs, e.g. in Free-Wilson analysis we are comparing certain molecular fragments, which means we are assuming they can interact similarly with the receptor or their atoms can be covered.[48] Eventually, QSAR implies a comparison, which means that molecular objects must be arranged in a certain *orientation* during such an operation. Most often a user must define this orientation. In this context an additional problem appears, namely, what is the relevance between the molecular superimposition modes for an individual molecule and a template and molecular recognition phenomena. In typical QSAR this question cannot be answered, but some novel approaches address such issues, as will be discussed below.

Although it is convenient to avoid superimposition this operation is extremely important for the results. In other words, if we get rid of this operation we can lose or modify information analyzed. Therefore, in the majority of 3D-QSAR we are controlling superimposition. In traditional approaches two possible responses (yes or no) answer the question of superimposition likelihood. We can improve model robustness by including some tolerance into the molecular superimposition. This can be interpreted as a fuzzy logic approach that forces a possibility for superimposition even if formally atoms cannot be covered providing a third additional answer, namely *atom congruence is acceptable*. Several improvements in the structure overlay have appeared that allow for more flexible or sophisticated superimposition.[49] In particular, neural networks have been used for flexible superimposition in Compass,[50] CoMSA,[51,52] or SOM-4D-QSAR.[53] Below we discuss the application of self-organizing neural network for the construction of fuzzy molecular representations that enable a control of the tolerance of molecular superimposition.[54]

Self-Organizing Neural Networks for Robust Superimposition. A transformation of 3D objects, e.g. the molecular surfaces, to 2D images always needs some projection or data transformation in order to reduce the dimension of the data to be visualized. Standard projections usually deform the topology of the objects to be transformed. In the early 1990s Zupan and Gasteiger developed a method for the Kohonen topographic mapping[55] of the molecular electrostatic potential to transform the 3D molecular surface data into a form of the 2D map.[56] For this transformation 3D coordinates sampled from the molecular surface are input directly to the competitive Kohonen neurons. Then, each output neuron of the 2D map is colored by the mean electrostatic potential value of the points projected from the 3D surface into this neuron. The preservation of the surface topology is among the most important issues of the transformation that is originally highlighted. Distinguishing between active and inactive molecules using Kohonen patterns was one of the first applications suggested.[57,58] Although there are some problems concerning a precise comparison of a pair of such maps, the method was used also for a so-called bioisosteric design.[59−61] However, if a single network is used for the processing of two different molecules in so-called *comparative mapping*, we obtain a kind of a precise superimposition tool.[58−62] The performance of such architecture in molecular design is thoroughly described in previous publications[45,51−54,58,59,61−64] and will not be further discussed here. The ability to generate fuzzy molecular surface representations which is of the importance for the explanation of the possible improvements in QSAR robustness is illustrated in Figure 5.

Comparative mapping if coupled with PLS analysis provided us a 3D-QSAR method, the Comparative Molecular Surface Analysis (CoMSA),[51] that is analogous to CoMFA but bases on the comparison of the surface sectors. The performance of CoMSA measured by single q^2 or SDEP parameters is generally better than CoMFA.[65,66] Figure 6 analyzes the robustness of CoMFA and CoMSA modeling of the CBG steroid activity by the SVM scheme.[35] It is clear that CoMSA gives higher q^2 and lower SDEP values. More recently Hasegawa improved the performance of CoMSA, additionally improving the method by coupling 3-way PLS,[67,68] and the non-neural CoMSA version has been developed.[52]

Data Reduction as a Robust Answer for Molecular Similarity Noise. Molecules are full of similarities, and it is not quite clear if in QSAR we are capable of extracting from the molecular structure only these aspects of similarity that is important for a certain activity. In fact, we lack the systematic investigations of this problem in QSAR. In classical QSAR different models can be constructed for a given molecular series, which of course brings a problem of model interpretation. For a discussion of this problem compare ref 69. It is observed in 3D-QSAR, namely in CoMFA, that different molecular descriptors give the final models of a similar statistical performance.[34] This indicates an effect that can be called molecular similarity noise, i.e., a fact that different analyses reveal different aspects of molecular correspondences that are however intercorrelated. Of course, similarity noise is unfavorable in 3D-QSAR because we can find molecular areas that are only coincidentally correlated with the activity and not the areas of the specific ligand−receptor interactions.

The CoMFA variables (describing individual molecular fields) included into a final model are weighted by the PLS analysis. Therefore, a final model contains information on the whole molecular field. We can indicate such an approach as the analysis of aggregated molecular similarity. If we were

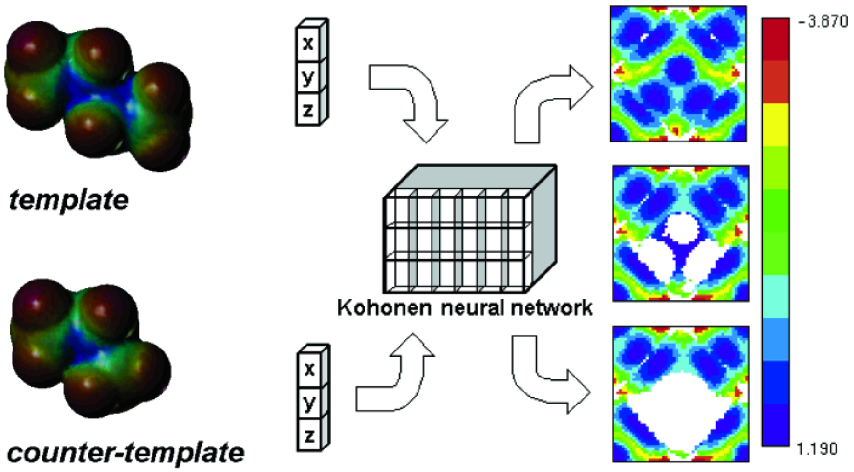


Figure 5. Fuzzy molecular surface representations by the comparative mapping of the butane and propane molecules. The molecules can be superimposed (no-empty, white, neurons are observed on two-dimensional propane map) or cannot be superimposed if the comparison is less tolerant (a large area of white neurons is observed on a two-dimensional propane map). Adapted from ref 54.

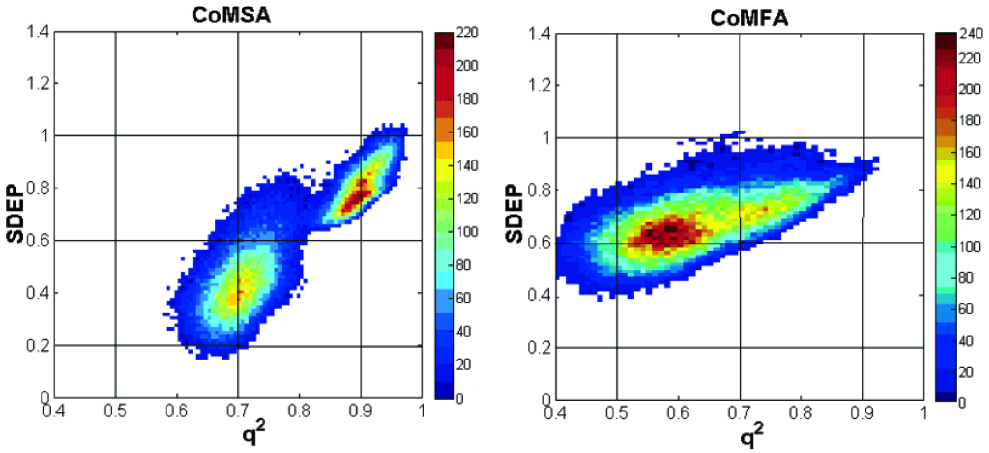


Figure 6. The SMV schemes for the validation of the CoMFA and CoMSA modeling of the CBG steroid series. Adapted from ref 35.

however capable of the indication of some among initial variables, then we could shift from a molecular similarity model to a pharmacophore based model. Generally, PLS only rarely combines with data elimination because this method works by *extracting* from each variable a right share to contribute for a total value.[29] However, CoMFA modeling coupled with data elimination can provide better results then the traditional CoMFA.[70,71] We have also shown that PLS combined with variable elimination, e.g. UVE or modified UVE versions, can be a powerful tool significantly improving both the predictive power of the CoMSA model and its illustrative ability. This in turn can bring an increased understanding of the molecular basis for the compounds biological interactions, e.g. steroid aromatase inhibitors.[72] Illustrative ability is especially worth mentioning in the aspect of pharmacophore mapping.[73] For better understanding Figure 7 shows the interaction contour plots for the series of hypolipidemic asarones.[74] Unlike CoMFA plots that are identical for all molecules, CoMSA provides a different illustration for each molecule. The incongruencies of molecular surfaces in individual molecules result in much more clear contour plots that are much easier for the interpretation. Thus, for example, we can observe that a carbonyl function is unfavorable for the activity of asarones.

Robust Answer for Conformational Noise. Biological response results from a certain atomic configuration (conformation). Since generally a variety of such molecular representations are available for a single molecule, this generates an effect that can be described by the term conformational noise. This raises the question if various conformational modes can produce statistically valid QSAR. 4D-QSAR addresses this issue by the investigations into the conformational space of molecules. In Hopfinger's 4D-QSAR molecular dynamic simulations provide *conformational ensemble profile* describing each molecule. Then, descriptors defining the pattern in which atoms occupy volume sectors are calculated.[75−80] Alternatively, a self-organizing neural network can be used for the generation of molecular volumes in SOM-4D-QSAR.[81] 4D-QSAR demands the variable elimination or selection step mounted as the integral filtering unit. Technically, Hopfinger's method uses the genetic algorithm for this purpose. In SOM-4D-QSAR we applied IVE-PLS.[52,72] 4D-QSAR can significantly improve model robustness even for such rigid molecules as

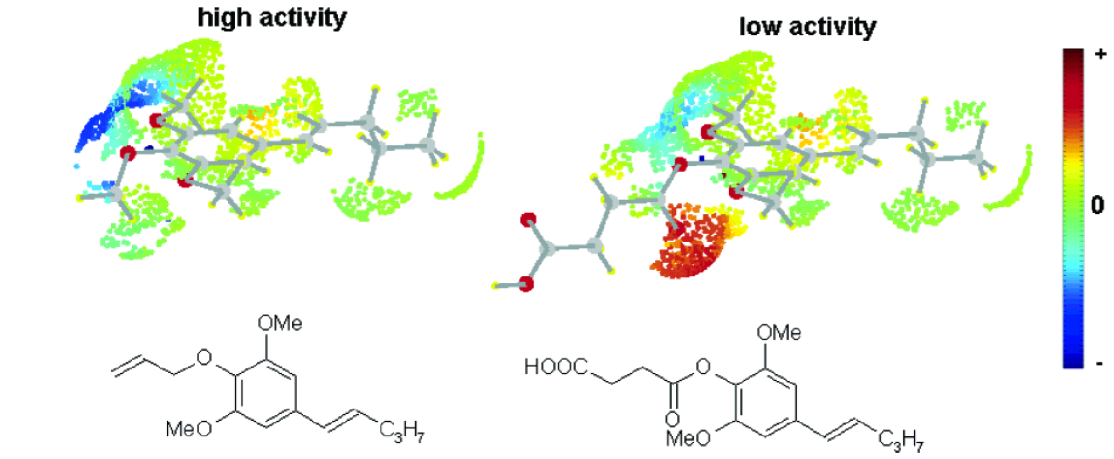


Figure 7. CoMSA interaction contour plots indicating the areas of the positive (activity decrease) and negative (activity increase) contribution for the activity in some arbitrarily selected high and low activity asarones. Adapted from ref 74.

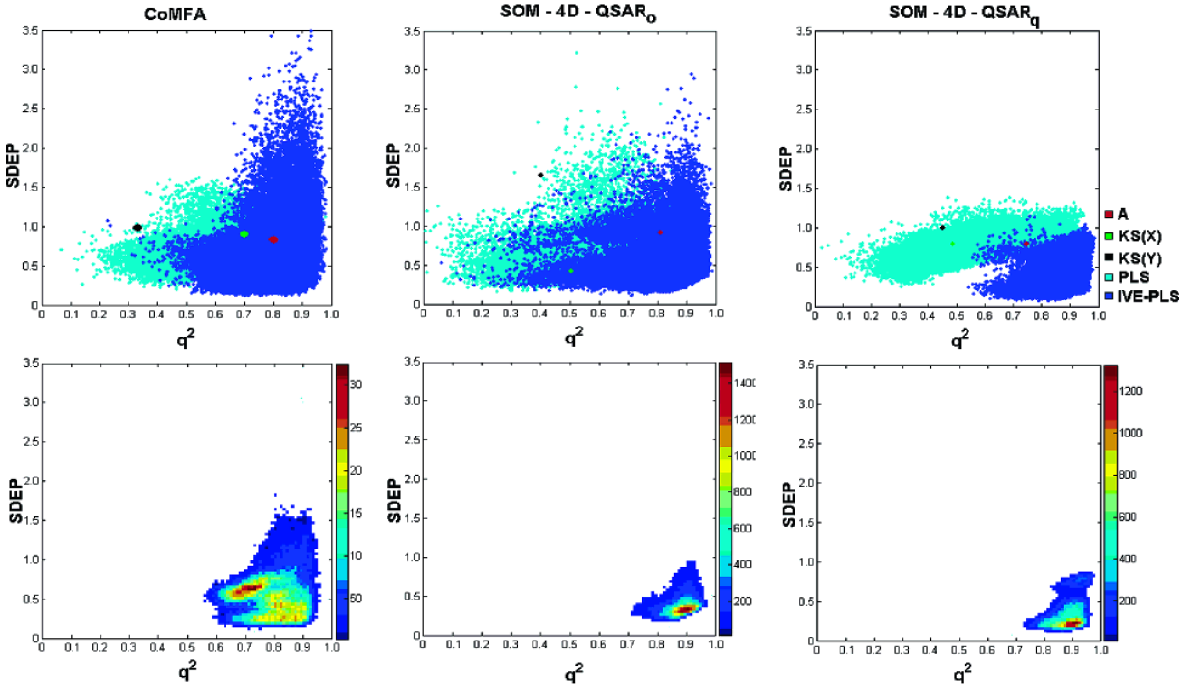


Figure 8. The SMV schemes for the estimation of the robustness of the CoMFA (a), SOM-4D-QSAR with occupancy descriptors (b) and SOM-4D-QSAR with charge descriptors (c) modeling of the CBG steroid series. Standard plots (upper line) illustrate the difference in the PLS and IVE-PLS modeling. Single points within the maps presented in the upper line indicate the following: A − a single validation usually reported in the literature, i.e., the model calculated for molecules 1−21 (training set) is cross-predicted to molecules 22−31 (test set); KS(X) is a single validation for the training and set sampled by the Kennard Stone protocol using the molecular descriptor (X) data; and KS(Y) is a single validation for the training and set sampled by the Kennard Stone protocol using the activity (Y) data. The density maps for the IVE-PLS models (bottom) are shown to illustrate the real distribution of the models.

steroid CBG series. This is illustrated in Figure 7 that shows SMV profiles for the CoMFA and SOM-4D-QSAR with occupancy and charge descriptors, respectively. Clearly, the latter model is also the most robust one. We have shown recently that SOM-4D-QSAR coupled with IVE-PLS is capable of the proper illustration of the HEPT inhibitor interaction with HIV-1 reverse transcriptase.[82] This not only gave a high quality model but also (as the only QSAR model reported) indicated a conformational mode that was consistent with real interactions determined by X-ray analysis.[83,84]

Robust Answer for Molecular Recognition Noise. QSAR is usually considered as a method that analyzes the data in a receptor-independent mode. This is not fully true, because biological activity data obviously depend on ligand−receptor interactions, but in fact generally a single number i.e., the activity value, accounts for the whole receptor. This forms a large imbalance in comparison to a data number describing a ligand molecule. Each molecule within the series analyzed in QSAR can interact within the receptor space receiving a specific ligand−receptor orientation. In a receptor

independent approach we are assuming that an orientation optimized during the molecular superimposition step describes also a relative orientation during molecular recognition phenomena. If the real ligand−receptor orientation differentiates individual molecules and does not comply with that optimized during the superimposition step, then a clear discrepancy between modeled and real data appears. Moreover, additional specific effects can be stimulated by ligand−receptor interactions during binding. This generates an important noise that can be described by the term *molecular recognition noise*. Therefore, an important improvement can be achieved by the incorporation of the receptor structure data into QSAR modeling. Thus, binding affinities can be calculated in silico for a series of molecules and a receptor structure and then modeled into an equation. The COMBINE is an example of such a method[85,86] that is developed for the calculation of ligand−receptor binding energies. The analysis of the HIV-1 RT inhibitor series is one of the latest applications of this method.[87] Alternatively, ligand−receptor data can also be used in QSAR modeling.[88,89]

CONCLUSIONS

QSAR is a term describing a variety of approaches that are of substantial interest for chemistry. This method can be defined as indirect molecular design by the iterative sampling of the chemical compounds space to optimize a certain property and thus indirectly design the molecular structure having this property. However, property production and modeling the interactions of chemical molecules in biological systems provides highly noisy data, which makes predictions a roulette risk. In this paper we briefly review the origins for this noise, particularly in multidimensional QSAR. This was classified as the data, superimposition, molecular similarity, conformational, and molecular recognition noise. We also indicated possible robust answers that can improve modeling and predictive ability of QSAR, especially self-organizing mapping of the molecular objects, in particular, the molecular surfaces, a method that was brought into chemistry by Gasteiger and Zupan.

REFERENCES AND NOTES

(1) Kolb, H.; Finn, G.; Sharpless, B. Click chemistry: Diverse chemical function from a few good reactions. *Angew. Chem., Int. Ed.* **2001**, *40*, 2004−2021.

(2) Mullin, R. Recalibrating the clinic. High-tech tools and streamlined business processes are making their way to the far reaches of the drug development pipeline. *C&EN* **2005**, *83*, 29−39.

(3) De Julian-Ortiz, J. Virtual Darwinian drug design: QSAR inverse problem. *Comb. Chem. High Throughput Screening* **2000**, *4*, 295−310.

(4) Hansch, C.; Leo, A. *Exploring QSAR: Fundamentals and applications in chemistry and biology*; American Chemical Society: Washington, DC, 1995.

(5) Cox, H.; Heaney, K. Approaches to robustness, *J. Acoust. Soc. Am.* **2003**, *113*, 2262−2262.

(6) Lipinski, A. Drug-like properties and the cause of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2004**, *44*, 235−249.

(7) Hann, M.; Oprea, T. Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* **2004**, *8*, 255−263.

(8) Oprea, T. Current trends in lead discovery. Are we looking for the appropriate properties? *J. Comput.-Aided Mol. Des.* **2002**, *16*, 325−334.

(9) Hansch, C.; Hoekman, D.; Leo, A.; Weininger, D.; Selassie, C. Chembioinformatics: Comparative QSAR at the interface between chemistry and biology. *Chem. Rev.* **2002**, *102*, 783−812.

(10) Hodgson, J. ADMET − turning chemicals into drugs. *Nat. Biotechnol.* **2001**, *19*, 722−726.

(11) Oprea, T. 3D-QSAR modeling in drug design. In *Computational Medicinal Chemistry and Drug Discovery*; Tolleneare, J., De Winter, H.; Langenaeker; W., Bultinck, P., Eds.; Marcel Dekker: New York, 2004.

(12) Kubinyi, H. Quantitative structure−activity relationships. The bilinear model, a new model for nonlinear dependence of biological activity on hydrophobic character. *J. Med. Chem.* **1977**, *20*, 625−629.

(13) Anzali, S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J.; Sadowski, J.; Teckentrup, A.; Wagener, M. The use of self-organizing neural networks in drug design. Kubinyi, H., Folkers, G., Martin, Y., Eds.; Kluwer: Dordrecht, The Netherlands, 1998.

(14) Xu, L.; Zhang, W. Comparison of different methods for variable selection, *Anal. Chim. Acta* **2001**, *446*, 477−483.

(15) Leardi, R. Genetic algorithms in chemometrics and chemistry: A review. *J. Chemom.* **2001**, *15*, 559−569.

(16) Saxena, K.; Prathipati, P. Comparison of MLR, PLS and GA-MLR in QSAR. *SAR QSAR Environ. Res.* **2003**, *14*, 433−445.

(17) Vapnik, V. N. *The nature of statistical learning theory*; Verlag Springer: New York, 1999.

(18) Furey, T.; Cristianini, N.; Duffy, N.; Bednarski, D.; Schummer, M.; Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **2000**, *16*, 906−914.

(19) Norinder, U. Support vector machine models in drug design: application to drug transport processes and QSAR using simplex optimizations and variable selection. *Neurocompiuting* **2003**, *55*, 337−346.

(20) Demiriz, A.; Bennet, K.; Breneman, C.; Embrechts, M. Support vector machine regression in chemometrics. *Comput. Sci. Stat.* **2001**, *33*, 289−296.

(21) Corne, D. W.; Martin A. C. Artificial intelligence in bioinformatics. *Comput. Chem.* **2002**, *26*, 1−3.

(22) David, V.; Sanchez, A. Advanced support vector machine and kernel methods. *Neurocomputing* **2003**, *55*, 5−20.

(23) Burbidge, R.; Trotter, M.; Buxton, B. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.

(24) Varmuza, K. Multivariate data analysis in chemistry. In *Handbook of chemoinformatics*; Wiley VCH: Verlag: Weinheim, 2003.

(25) Esbensen, S.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37−52.

(26) Geladi, P.; Kowalski, B. Partial least squares: A tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

(27) Helland, I. Some theoretical aspects of partial least squares regression. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 97−107.

(28) Good, P. *Resampling methods: A practical guide to data analysis*; Birkhauser: 1999.

(29) Wold, S.; Sjöström, M.; Eriksson, L. In *The Encyclopedia of Computational Chemistry*; Wiley and Sons: Chichester, U.K., 1999.

(30) Wakeling, N.; Morris, J. A test of significance for partial squares regression. *J. Chemom.* **1993**, *7*, 291−304.

(31) Clark, M.; Crammer III, R. The probability of chance correlation using partial least squares (PLS). *Quant. Struct.−Act. Relat.* **1993**, *12*, 137−145.

(32) Tropsha, A.; Gramatica, P.; Gombar, K. The importance on being earnest: Validation is the absolute essential for successful application and interpretation of QSAR models. *QSAR* **2003**, *22*, 69−77.

(33) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

(34) Daweyko, A. 3D-QSAR illusions. *J. Comput.-Aided Mol. Des*. **2004**, *18*, 587−596.

(35) Polanski, J.; Gieleciak, R.; Bak, A. Probability issues in molecular design: Predictive and modeling ability in 3D-QSAR schemes. *Comb. Chem. High Throughput Screening* **2004**, *7*, 793−807.

(36) Clark, R. Boosted leave-many-out cross-validation: The effect of training and test set diversity on PLS statistics. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 265−275.

(37) Sheridan, R.; Feuston, B.; Maiorov, V.; Kearsley, S. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912−1928.

(38) Cramer III, R.; Patterson, D.; Bunce, J. Comparative molecular field analysis (CoMFA). *J. Am. Chem. Soc.* **1998**, *110*, 5959−5967.

(39) Kubinyi, H. Comparative molecular field analysis (CoMFA). In *Handbook of Chemoinformatics. From data to knowledge*; Gasteiger, J., Ed.; Wiley VCH: BRD, Weinheim, 2003.

(40) Melani, F.; Gratteri, P.; Adamo, M.; Bonaccini, C. Field interaction and geometrical overlap: A new simplex and experimental design based computational procedure for superposing small ligand molecules. *J. Med. Chem.* **2003**, *46*, 1359−1371.

(41) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215−232.

(42) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic Ah receptor activity by neural networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.
(43) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley VCH: BRD, Weinheim, 1999.
(44) Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **1994**, *37*, 4130−4136.
(45) Polanski, J. The receptor like neural network for modeling corticosteroid and testosterone binding globulins. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 553−561.
(46) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRID-independent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.
(47) Silverman, B.; Platt, D. Comparative molecular field moment analysis (CoMMA). *J. Med. Chem.* **1996**, *39*, 2129−2140.
(48) Free, S.; Wilson, J. A mathematical contribution to structure -activity studies. *J. Med. Chem.* **1964**, *7*, 395−399.
(49) Korhonen, S. P.; Tuppurainen, K.; Laatikainen, R.; Peräkyla, M. FLUFF-BALL A template-based grid-independent superposition and QSAR technique: Validation using a benchmark steroid data set. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1780−1793.
(50) Jain, A.; Koile, K.; Champan, D. Compass: Predicting biological activities from molecular surface properties. Performance comparison on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315−2327.
(51) Polanski, J.; Walczak, B. The comparative molecular surface analysis (CoMSA): A novel tool for molecular design. *Comput. Chem.* **2000**, *24*, 615−625.
(52) Polanski, J.; Gieleciak, R.; Magdziarz, T.; Bak, A. The grid formalism for the comparative molecular surface analysis: Application to the CoMFA benchmark steroids, azo dyes and HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1423−1435.
(53) Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. Self-organizing neural network for modeling robust 3D and 4D QSAR: Application to dihydrofolate reductase inhibitors. *Molecules* **2004**, *9*, 1148−1159.
(54) Polanski, J. Molecular Shape Analysis. In *Handbook of Chemoinformatics. From data to knowledge*; Gasteiger, J., Ed.; Wiley VCH: BRD, Weinheim, 2003.
(55) Kohonen, T. *Self-organizing and associate memory*, 3rd ed.; Springer; Berlin, 1989.
(56) Gasteiger, J.; Li, X.; Rudolph, C.; Sadowski, J.; Zupan, J. Representation of molecular electrostatic potentials by topological feature maps. *J. Am. Chem. Soc.* **1994**, *116*, 4608−4620.
(57) Gasteiger, J.; Zupan, J. Neural networks in chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503−512.
(58) Polanski, J.; Gasteiger, J.; Wagener, M.; Sadowski, J. The comparison of molecular surfaces by neural networks and its application to quantitative structure activity studies. *Quant. Struct.−Act. Relat.* **1998**, *17*, 27−36.
(59) Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. The comparison of geometric and electronic properties of molecular surfaces by neural networks: Applications to the analysis of corticosteroid binding globulin activity of steroids. *J. Comput-Aided Mol. Des.* **1996**, *10*, 521−534.
(60) Anzali, S.; Maderski, W.; Osswald, M.; Dorsch, D., Endothelin antagonists: Search for surrogates of methylendioxyphenyl by means of a Kohonen neural network. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 11−16.
(61) Polanski, J.; Gasteiger, J.; Jarzembek, K. Self-organizing neural networks for screening and development of novel artificial sweetener candidates. *Comb. Chem. High Throughput Screening* **2000**, *3*, 481−495.
(62) Barlow, T. Self-organizing maps and molecular similarity. *J. Mol. Graphics* **1995**, *13*, 24−27.
(63) Polanski, J.; Gasteiger, J. The comparison of molecular surface by assembly of self-organizing neural network, In proceedings of the III-th International Conference "Computers in Chemistry '94", Technical University of Wroclaw, Wroclaw, Poland, 1994, p 88.
(64) Livingstone, D.; Manallack, D. Neural networks in 3D QSAR. *QSAR Comb. Sci.* **2003**, *22*, 510−518.
(65) Polanski, J.; Gieleciak, R.; Bak, A.; Jarzembek, K.; Wyszomirski, M. The comparative molecular surface analysis (CoMSA). A novel efficient technique for drug design. *Acta Pol. Pharm.* **2002**, *59*, 459−461.
(66) Polanski, J.; Gieleciak, R.; Bak, A. The comparative molecular surface analysis (CoMSA) − A nongrid 3D QSAR method by a coupled neural network and PLS system: Predicting pK_a values of benzoic and alkanoic acids. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 184−191.
(67) Hasegawa, K.; Morikami, K.; Shiratori, Y.; Ohtsuka, T.; Aoki, Y.; Shimma, N. 3D-QSAR study of antifungal N-myristoyltra transferase inhibitors by comparative molecular surface analysis. *Chemom. Intell. Lab. Syst.* **2003**, *69*, 51−59.
(68) Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatsu, K. New molecular surface-based 3D-QSAR metod using Kohonen neural network and 3-way PLS. *Comput. Chem.* **2002**, *26*, 583−589.
(69) Wermuth, C. The impact of QSAR and CADD methods in drug discovery In *Rational approach to drug design*; Höltji, H., Sippl, W., Eds.; Prous Science: Barcelona, 2001.
(70) Cho, S.; Tropsha, A. Cross-validated r^2-guided region selection for comparative molecular field analysis: A simple method to achieve consistent results. *J. Med. Chem.* **1995**, *38*, 1060−1066.
(71) Cho, S.; Tropsha, A.; Suffnes, M.; Cheng, Y.; Lee, K. Antitumor agents. 163. Three-dimensional quantitative structure activity relationship study of 4′-O-dimethylepipodophyllotoxin analogues using the modified CoMFA/q^2-GRS approach. *J. Med. Chem.* **1996**, *39*, 1383−1395.
(72) Polanski, J.; Gieleciak, R. The comparative molecular surface analysis (CoMSA) with modified uninformative variable elimination PLS (UVE-PLS) method: Application to the steroids binding the aromatase enzyme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 656−666.
(73) Polanski, J. Self-organizing neural networks for pharmacophore mapping. *Adv. Drug Deliv. Rev.* **2003**, *55*, 1149−1162.
(74) Gieleciak, R.; Magdziarz, T.; Bak, A.; Polanski, J. Modeling robust QSAR 1: Coding molecules in 3D QSAR − from a point to surface sectors and molecular volumes. *J. Chem. Inf. Model.* **2005,** *45***,** 1447−1455.
(75) Hopfinger, A.; Wang, S.; Tokarski, J.; Jin, B.; Albuquerque, M.; Madhav, P.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.
(76) Albuquerque, M.; Hopfinger, A.; Barreiro, E.; De Alencastro, R. Four-dimensional quantitative structure−activity relationship analysis of a series of interphenylene 7-oxabicycloheptane oxazole thromboxane A_2 receptor antagonists. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 925−938.
(77) Santos-Filho, O.; Hopfinger, A. A search for sources of drug resistance by the 4D-QSAR analysis of a set of antimalarial dihydrofolate reductase inhibitors. *J. Comput-Aided Mol. Des.* **2001**, *15*, 1−12.
(78) Ravi, M.; Hopfinger, A.; Hormann, R.; Dinan, L. 4D-QSAR analysis of a set of ecdysteroids and a comparison to CoMFA modeling. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1587−1604.
(79) Krasowski, M.; Hong, X.; Hopfinger, A.; Harrison, N. 4D-QSAR analysis of a set of propofol analogues: Mapping binding sites for an anesthetic on the $GABA_A$ receptor. *J. Med. Chem.* **2002**, *45*, 3210−3221.
(80) Hong, X.; Hopfinger, A. 3D-pharmacophores of flavonoid binding at the benzodiazepine $GABA_A$ receptor site using 4D-QSAR analysis. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 324−336.
(81) Polanski, J.; Bak, A. Modeling steric and electronic effects in 3D- and 4D-QSAR schemes: Predicting benzoic pK_a values and steroic CBG binding affinities. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2081−2092.
(82) Bak, A.; Polanski, J. The 4D-QSAR study on anti-HIV HEPT analogues. *Bioorg. Med. Chem.* **2006**, *14*, 273−279.
(83) Kireev, D.; Chrétien, J.; Grierson, D.; Monneret, C. A 3D-QSAR study of a series of HEPT analogues: The influence of conformational mobility on HIV-1 reverse transcriptase inhibition. *J. Med. Chem.* **1997**, *40*, 4257−4264.
(84) Jalali-Heravi, M.; Parastar, F. Use of artificial neural networks in a QSAR study of anti-HIV activity for a large group of HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **2000,** *40*, 147−154.
(85) Murcia, M.; Ortiz, A. Virtual screening with flexible docking and COMBINE-based models. Application to a series of factor Xa inhibitors. *J. Med. Chem.* **2004**, *47*, 805−820.
(86) Wang, T.; Wade, R. Comparative binding energy (COMBINE) analysis of OppA-peptide complexes to relate structure to binding thermodynamics. *J. Med. Chem.* **2002**, *45*, 4828−4837.
(87) Rodriguez-Barrios, F.; Gago, F. Chemometrical identification of mutations in HIV-1 reverse transcriptase conferring resistance or enhanced sensitivity to arylsulfonylbenzonitriles. *J. Am. Chem. Soc.* **2004**, *126*, 2718−2719.
(88) Rondeau, J. M.; Schreuder, H. Protein Crystallography and Drug Discovery, In *The Practice of Medicinal Chemistry*; Wermuth, C. G., Ed.; Academic Press: 2003; pp 417−443.
(89) Sippl, W.; Contreras, M.; Parrot, I.; Rival, M.; Wermuth, G. Structure-based 3D-QSAR and design of novel acetylcholinesterase inhibitors. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 395−410.

CI050314B

Modeling Robust QSAR. 1. Coding Molecules in 3D-QSAR − from a Point to Surface Sectors and Molecular Volumes

Rafal Gieleciak, Tomasz Magdziarz, Andrzej Bak, and Jaroslaw Polanski*

Department of Organic Chemistry, Institute of Chemistry, University of Silesia, PL-40-006 Katowice, Poland

Received April 26, 2005

Shape analysis is a powerful tool in chemistry and drug design. In the current work, we compare the results of CoMFA and Comparative Molecular Surface Analysis (CoMSA), the 3D-QSAR method, for a series of hypolipidemic and antiplatelet asarones and antifungal N-myristoyltransferase inhibitors. In this publication we show that a sector CoMSA formalism enables an analysis of the biological activity that is more directly related to the molecular shape and individual molecular functionalities than the traditional uniform and directionless CoMFA field. Iterative Variable Elimination allowed us to identify the potential pharmacophoric sites. We modeled QSARs for both series and demonstrate that sector-based molecular descriptors give very predictive models and allow one to generate a spatial interpretation of the QSAR models. In particular, we identified the central aromatic ring and carbonyl functions as the moieties determining the activity of the asarones series, while the pattern of substitution of the aromatic ring determines the activity of N-myristoyltransferase inhibitors.

INTRODUCTION

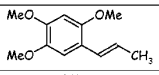
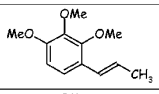
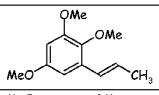
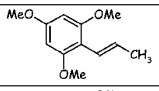
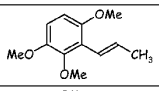
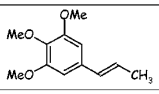
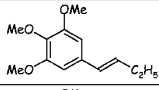
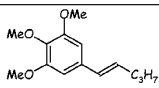
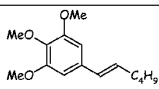
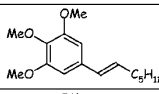
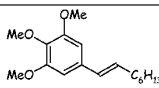
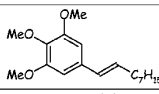
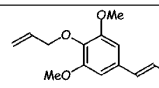
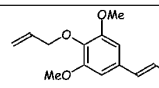
It may appear to be a paradox, but *the most fundamental and lasting objective of* (chemical) *synthesis is not the* production of new compounds but the production of properties.[1] Molecular design is a computational tool for screening virtual chemical compound space in a search for novel properties, and QSAR should function like a dictionary between molecular structures and properties. This clearly makes it an essential and irreplaceable method in molecular design. However, more and more sophisticated tools are needed for the efficient and robust transformation of molecular structure space into compound property space. A variety of issues decide the efficiency of the QSAR methods,[2] and their practical importance for drug design is still controversial.[3] Although it is common to think of QSAR as a drug design method, in fact, this technique is more an a posteriori data analysis process than real design. First, it is not the same to start from a molecular structure and attempt to predict its properties as to begin from a property and find the molecule having this particular property. Second, the term 'design' anticipates the extrapolation of the data to novel objects. However, the fact that a very minute molecular modification can evoke a substantial activity change makes such extrapolation extremely risky. Molecular superposition and conformational flexibility present further problems for QSAR modeling. In this context, a number of possible molecular arrangements and configurations can be generated before multidimensional QSAR modeling. This makes QSAR a highly data dependent operation and introduces substantial noise into QSAR results. Can we decrease data dependency in QSAR, making it less sensitive to the variation in inputs using novel, more robust systems. Different superimposition rules provide completely different activity contour plots in CoMFA.[4] Recently, several improvements in structure overlay have appeared that allow for more flexible or sophisticated superimposition.[5] Hopfinger's 4D-QSAR investigating molecular conformational space has been supplemented by architecture that uses a fuzzy self-organizing neural neuron.[6−8] Data handling can improve QSAR robustness, and the application of new computational methods including neural networks, data elimination, genetic algorithms, and novel model validation schemes have often been reported in this field. Essentially, the efficiency of data handling largely depends on the descriptors that are used for the characterization of the molecular objects analyzed. Thus, the method that is used for coding molecules in 3D- or 4D-QSAR is an important factor that does influence final model robustness. In the CoMFA-like fields a molecule is represented by a set of points determined in space by a 3D grid.[9] Different smooth and box CoMFA-like fields have recently been thoroughly tested.[10] A surface can serve as the base for the molecule description in several methods, e.g., Compass, CoRSA, SURFCOMP, or CoMSA, which are based on sampling points on the molecular surface or near such a surface.[11−14] Various algorithms have been developed for the comparison of the surface sectors. Thus, the lattice generated in CoRSA on the molecular surface defines the nodes that are further compared for the series of individual molecules. Alternatively, to compare surface sectors for a series of molecules, molecular volumes can be defined in analyzed molecules. Different algorithms can be used to generate such volumes, which can take the form of a rectangular cube (s-CoMSA)[15] or a sphere generated by self-organizing neural networks (SOM-CoMSA)[16−22] or a supervised neural network (Compass).[11] Similarly, in Hopfinger's 4D-QSAR a molecule is coded openly by the descriptors defining the pattern in which atoms occupy volume sectors.[23] Alternatively, a self-organizing neural network can be used for the generation of molecular volumes in SOM-4D-QSAR.[24]

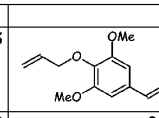
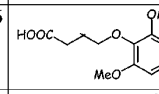
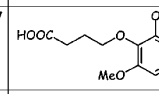
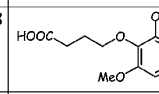
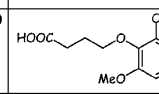
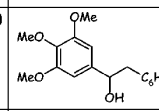
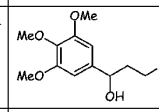
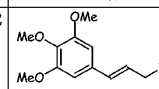
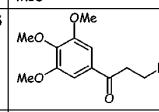
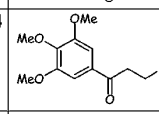
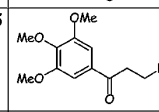
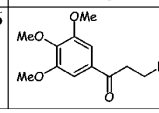
In the present work we investigate the influence of the way in which molecules are coded on the efficiency and

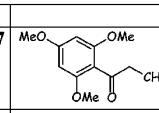
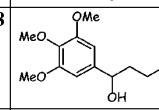
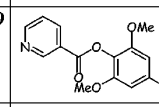
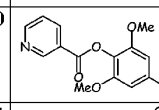
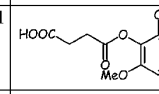
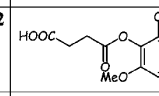
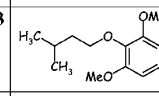
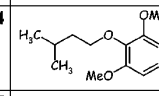
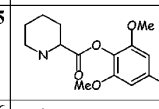
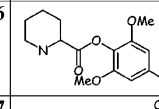
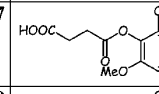
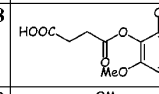
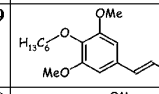
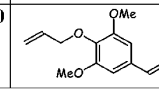
* Corresponding author e-mail: polanski@us.edu.pl.

10.1021/ci0501488 CCC: $30.25
Published on Web 08/31/2005

Table 1. α-Asarone Compounds and Their Atherogenic Index − $I_{TG/LDL}$

		$I_{TG/LDL}$
a1		2.10
a2		1.27
a3		1.56
a4		1.34
a5		2.35
a6		1.31
a7		0.36
a8		0.45
a9		0.16
a10		0.27
a11		0.13
a12		0.14
a13		0.47
a14		0.15

		$I_{TG/LDL}$
a15		0.10
a16		0.45
a17		0.22
a18		0.13
a19		0.85
a20		0.77
a21		1.74
a22		2.05
a23		2.00
a24		1.38
a25		1.48
a26		1.71

		$I_{TG/LDL}$
a27		1.90
a28		1.26
a29		3.15
a30		1.52
a31		1.79
a32		3.28
a33		0.45
a34		0.56
a35		1.98
a36		2.58
a37		1.92
a38		1.96
a39		1.28
a40		0.80

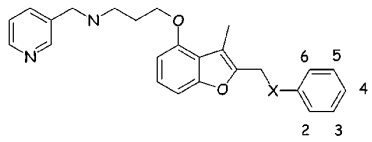
robustness of 3D-QSAR modeling. In particular, we compare the traditional molecular point field generated by the CoMFA and volume-based descriptors used by the CoMSA method. Molecules are rich in analogies, and different molecular descriptors can provide similar QSAR models. Although it is believed that the method used for the calculation of partial atomic charges highly influences resultant QSAR, Doweyko[4] showed that CoMFA provides models of a similar statistical quality independent of the calculation method used. In a context of QSAR, this problem introduces further noise, namely, *molecular similarity noise*. In our investigations we analyzed the application of the Iterative Variable Elimination[19] for the indication of the molecular areas that are specific to biological activity, which should reduce similarity noise and indicate possible pharmacophoric sites.

We investigated asarones **a1**−**a40** and benzofuranes **b1**−**b29**, the same compound series that have previously been investigated with 3D-QSARs.[25,26]

EXPERIMENTAL SECTION

Model Builders. All of the experimental data, i.e., **a1**−**a40** and **b1**−**b29**, were extracted from refs 25 and 26 and are given in Tables 1 and 2, respectively.

Table 2. NMT Inhibitory Activity of Benzofurans



no.	X	R2	R3	R4	R5	R6	log(1/IC$_{50}$)
b1	O	F	H	F	H	H	8.12
b2	O	H	CF_3	H	H	H	6.72
b3	O	H	H	H	H	H	7.14
b4	O	H	H	Cl	H	H	7.14
b5	S	H	H	H	H	H	6.21
b6	S	H	H	Cl	H	H	5.71
b7	O	F	H	H	H	H	8.08
b8	O	H	F	H	H	H	6.95
b9	O	F	H	H	H	F	7.47
b10	O	F	H	H	F	H	8.36
b11	O	F	F	H	H	H	8.44
b12	O	F	F	H	F	H	8.18
b13	O	F	H	F	F	H	8.03
b14	O	F	F	F	H	H	8.24
b15	O	F	F	H	H	F	7.48
b16	O	F	H	F	H	F	7.09
b17	O	F	F	F	F	F	5.85
b18	O			2-Py[a]			5.53
b19	O			3-Py[a]			7.24
b20	O			4-Py[a]			5.81
b21	O	CN	H	H	H	H	7.78
b22	O	H	CN	H	H	H	7.03
b23	S	H	H	F	H	H	5.78
b24	O	F	F	H	F	F	6.24
b25	O	F	H	Br	H	H	7.55
b26	O	H	Br	H	H	H	6.40
b27	O	H	H	Br	H	H	6.06
b28	O	2-Py[a]	Cl	H	H	H	6.49
b29	O	2-Py[a]	H	Cl	H	H	6.64

[a] Py − pyridine.

All of the molecules were superimposed prior to calculation of the molecular surfaces. The superimposition was performed by covering the central benzene ring of molecules (**a1**−**a40**) and the central benzofurane ring of molecules (**b1**−**b29**). We used the program Match3D for performing this operation.[27] Calculation of the molecular surface descriptors was based on molecular volumes (SOM-CoMSA and s-CoMSA).

The SOM-CoMSA formalism based on the self-organizing neural network has been previously described[15−22] and will not be detailed here. Hasegawa described a similar but slightly modified procedure.[28,29] For the calculation of shape s-CoMSA descriptors we used formalism similar to Hopfinger's 4D-QSAR grid coding system.[23] Thus, each 3D molecular representation is placed in its own virtual cubic grid, and the molecular surface is calculated, respectively. The electrostatic potential is calculated for points randomly sampled on the molecular surface, and a mean value of the electrostatic potential corresponding to the respective points found in each grid cell is used to describe this cell. Grid cells are unfolded into vectors, and vectors describing all molecules of the series are aligned into a matrix. Grid cells that are empty for all molecules in the series analyzed are eliminated, and the resulting matrix was used for further calculations using the PLS method.

Formally, the descriptors are defined as follows. Each molecule m is represented by a set of points sampled on the molecular surface $P_m(x, y, z, v)$ where x, y, and z are coordinates in three-dimensional space and v represents a surface property in a given point, e.g., v is an electrostatic potential value.

Let $P_{mi}(x, y, z, v)$ be a subset of $P_m(x, y, z, v)$ such that all its elements are included in sector i. This is satisfied when

$$x_k \geq x_{l_i} \wedge x_k < x_{h_i} \wedge$$

$$y_k \geq y_{l_i} \wedge y_k < y_{h_i} \wedge$$

$$z_k \geq z_{l_i} \wedge z_k < z_{h_i} \qquad (1)$$

where x_{l_i}, y_{l_i}, z_{l_i} and x_{h_i}, y_{h_i}, z_{h_i} are the elements of vectors l_i and h_i which define the highest and lowest sector borders, respectively, and where i indexes a sector and k indexes the points sampled within the individual sectors. P_{mi} is an empty set if no k satisfies condition (1).

The s-CoMSA calculation lies in the generation of matrix D_A of size $g \times n$ where g is a number of analyzed molecules and n is the total number of sectors, and where d_{mi} is given by

$$d_{mi} = \begin{cases} \dfrac{\sum_{k=1}^{k_{\max}} (v_{mi})_k}{k_{\max}}, & \text{when } k_{\max} \neq 0 \\ 0, & \text{when } k_{\max} = 0 \end{cases} \qquad (2)$$

and $k_{\max}$ is the number of points on the surface P of the molecule m enclosed in sector i.

Matrix D defined by eq 2 describes each molecule independently of all others. After 4D-QSAR nomenclature, we defined such a descriptor absolute occupancy, D_A.

PLS analysis: vectors obtained were processed by the PLS analysis with a leave-one-out cross-validation procedure. The PLS procedures were programmed within the MATLAB environment (MATLAB).[30]

A PLS model was constructed for the centered data, and its complexity was estimated based on the leave-one-out cross-validation (CV) procedure. In the leave-one-out CV one repeats the calibration m times, each time treating the ith left-out object as the prediction object. The dependent variable for each left-out object is calculated based on the model with one, two, three, etc. factors. The Root Mean Square Error of CV for the model with j factors is defined as

$$\mathrm{RMSECV}_j = \sqrt{\frac{\sum(\mathrm{obs} - \mathrm{pred}_j)^2}{m}} \qquad (3)$$

where *obs* denotes the assayed value and *pred* denotes the predicted value of the dependent variable. A model with k factors, for which RMSECV reaches a minimum, is considered as an optimal one.

For the construction of the individual model reported in this work, the optimal number of latent PLS variables was truncated not to exceed the value of 1−10, respectively, which is clearly indicated in the figures.

We used performance metrics that are widely accepted and used in CoMFA analyses, i.e., cross-validated q^2_{cv}

$$q_{cv}^2 = 1 - \frac{\sum(\text{obs} - \text{pred})^2}{\sum(\text{obs} - \text{mean(obs)})^2} \quad (4)$$

where *obs* is the assayed values, *pred* is the predicted values, mean is the mean value of obs, and the cross-validated standard error *s*

$$s = \sqrt{\frac{\sum(\text{obs} - \text{pred})^2}{m - k - 1}} \quad (5)$$

where m is the number of objects and k is the number of PLS factors in the model.

Before PLS analysis was performed, the descriptors were centered, and this operation was repeated for each cross-validation run.

The quality of external predictions was measured by the SDEP parameter

$$\text{SDEP} = \sqrt{\frac{\sum(\text{pred} - \text{obs})^2}{n}} \quad (6)$$

where *pred* is the predicted value, *obs* is the observed value, and n is the number of measurements.

Data Elimination. To identify the parts of the molecular surface that contribute the most to activity, we used a modified procedure of the PLS with Uninformative Variable Elimination (UVE-PLS), namely the Iterative Variable Elimination PLS (IVE-PLS) procedure.[19] The UVE algorithm was originally proposed by Centner et al.[31] as a possible improvement to the PLS models. The main idea of the method is to reduce the number of variables included in the final PLS model. The UVE algorithm is based on the analysis of the regression coefficients calculated by the PLS method. The PLS method allows the presentation of the relation between the **Y** answer and **X** predictors in the form of

$$\mathbf{Y} = \mathbf{X}b + e$$

where b is a vector of the regression coefficients and e is the vector of the errors.

Thus, the UVE algorithm analyzes the reliability of the mean(b)/s(b) ratio (where s(b) means standard deviation of b). Then, only the variables of the "relative" high mean(b)/s(b) ratio are included in the final PLS model. To estimate the cutoff level artificial random number noise is generated (the level of the noise is 10^{-10} of the original variable order) and included (as additional columns) in the matrix of the original variables. PLS analysis of such a matrix is performed, and the mean(b)/s(b) parameter is analyzed for each column. The highest absolute value, abs(mean(b)/s(b)), observed in the noisy column determines the cutoff level for the original variables.

Modified Uninformative Variable Elimination based on Iterative Leave-One-Out Cross-Validation (IVE-PLS). Below we describe a modified procedure for Uninformative Variable Elimination (UVE-PLS). Instead of a single step procedure, we used here an iterative algorithm based on the abs(mean(b)/s(b)) criterion to identify variables to be eliminated. To distinguish this procedure, we named this Modified Uninformative Variable Elimination with the iterative leave-one-out cross-validation (IVE-PLS). This procedure includes the following: 1. standard PLS analysis applied to analyze the matrices yielded from the s-CoMSA procedure with the leave-one-out cross-validation to estimate the performance of the PLS model (q^2), 2. elimination of the matrix column of the lowest abs(mean(b)/s(b)) value, 3. standard PLS analysis of the new matrix without the column eliminated in step 2, and 4. iterative repetition of the steps 1−3 to maximize the LOO CV q^2 parameter.

The UVE and IVE procedures were programmed within the MATLAB environment (MATLAB).[30] All MATLAB functions and m-scripts are available from the authors upon request.

RESULTS AND DISCUSSION

Separating the molecule into partitions of spatial regions of certain volumes, either filled or unfilled by atoms or groups of atoms, can provide an interesting method for calculation of the molecular descriptors for efficient QSAR modeling.[32] Hopfinger's 4D-QSAR method uses similar formalism for the description of the molecular conformational space. Thus, in Figure 1 we compare some of the methods using the sector formalism (CoMSA, receptor-like neuron network, 4D-QSAR) to the conventional CoMFA point formalism. We show below that such formalism not only increases the fuzziness of the molecular description[13] but also significantly influences the performance of the QSAR model generated. It also changes the visual pattern in which such a QSAR model can be presented.

Hypolipidemic and Antiplatelet Asarones. Excess food intake makes lipid metabolism disorders a common plague of contemporary society. The therapy of such disorders is based on drugs having hypolipidemic activity, such as fibric acid derivatives, the inhibitors of 3-hydroxymethyl-conzyme A (3-HMG Co-A) reductase, an enzyme involved in de novo sterols synthesis, probucol, lifibrol, and others.[33,34] α-Asarones are compounds having hypolipidemic activity[35,36] and have been investigated in many pharmacological,[37,38] toxicological,[39,40] and QSAR studies.[41,42] We have recently applied CoMSA analysis for the design of potential asarone drugs. Cross-validated q^2_{cv} values range from 0.49; $s = 0.66$ (CoMFA) to $q^2_{cv} = 0.69$; $s = 0.54$ (SOM-CoMSA). This demonstrates a correlation between the descriptors and the hypolipidemic activity of the asarone series.[43] In Figure 2, we analyzed a model performance during the IVE-PLS variable elimination. This procedure was performed in such a way that the optimal number of PLS latent components was always estimated. This number, however, was truncated not to exceed the values of 1−10, respectively. We conclude this section with a few general remarks. First, CoMFA and s-CoMSA provide initial models of similar q^2_{cv} performance of ca. $q^2_{cv} = 0.5$. The neural SOM-CoMSA technique allowed us to increase this value to ca. $q^2_{cv} = 0.7$, even without data elimination. IVE-PLS enabled an increase in model performance measured by q^2_{cv} to ca. 0.7 (CoMFA) or 0.8 (SOM-CoMFA); 0.9 (s-CoMSA). For both CoMSA methods, model quality clearly depends on the maximal number of the PLS latent variables that can be included in the model. In contrast, the q^2_{cv} value in IVE-PLS-CoMFA does not depend on this number. We would also like to stress

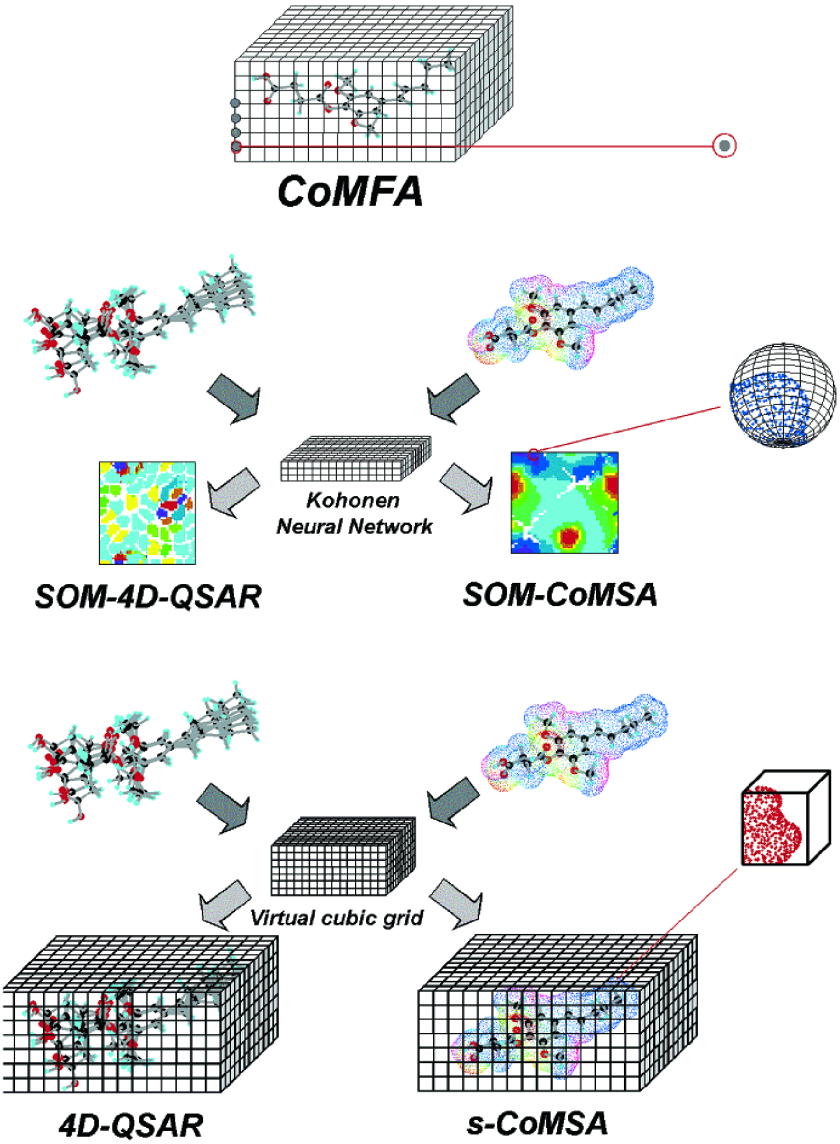


Figure 1. A schematic illustration comparing a uniform CoMFA grid (a sharply defined point defines molecular feature) with the neural SOM-CoMSA and SOM-4D-QSAR (a fuzzy SOM neuron defined defines molecular feature) and sector (Hopfinger's) 4D-QSAR and s-CoMSA (cubic sector defines molecular feature).

that external predictability does not change during IVE-PLS data elimination if tested by the SDEP parameter (results not shown). A second conclusion is that although the neural method provides a better starting model, data elimination is more efficient for the s-CoMSA. This indicates that the additional model improvement due to neural network fuzziness cannot be easily *extrapolated* to other models, and IVE-PLS does not enhance this extra q^2_{cv} gain.

Figure 3 compares the visualization of the compound's activity by the CoMFA and CoMSA methods. Thus, the uniform CoMFA field filtered by either a standard deviation value (Figure 3a) or further transformed by data elimination (IVE-PLS − Figure 3b) gives a uniform illustration for all molecules. Unlike CoMFA, both CoMSA methods explain the compounds' activity by the indication of the areas that can be easily differentiated for active and inactive molecules. This effect results from the dissimilarity of the molecular surfaces of the individual molecules. The respective color-coding allows us to identify the influence of the points sampled on the molecular surface by the combination of the electrostatic value and a value of the b weight in the PLS model. Points contributing to the activity on a level close to 0 (near 90% of the points sampled) were omitted. Such an illustration suggests the key pharmacophore for the asarones investigated. Thus, an area near the central aromatic ring substituted with alkoxyl functionality provides a negative contribution (blue-colored sections). Since lower values indicate higher activity, a *negative contribution* increases a compound's activity. A negatively charged carbonyl oxygen in the side chain generally causes lower activity, clearly decreasing the activity as illustrated by the yellow and red molecular surface areas. This rule can be proved by

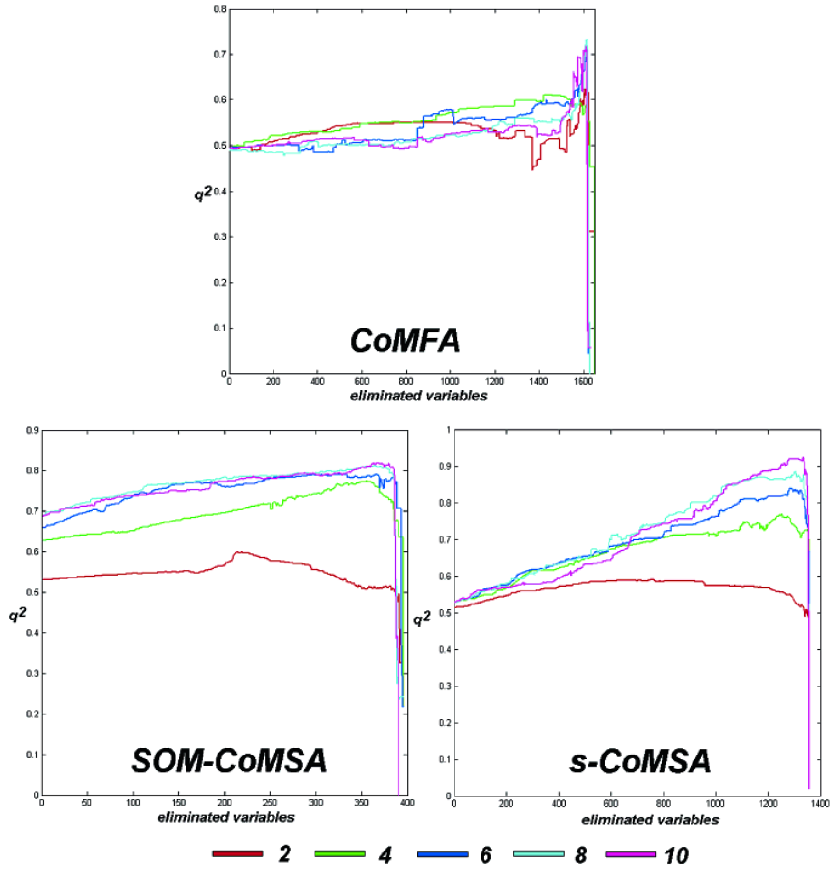


Figure 2. Profiles reporting the IVE-PLS modeling processes for the CoMFA and CoMSA for the asarones series, details in text.

examination of the structures given in Table 1. It is worth noticing that if a hydroxyl function replaces a carbonyl oxygen, e.g. compound **a28**, we do not observe an activity decrease similar to that affected by a carbonyl group, e.g., compound **a26**.

Antifungal N-Myristoyltransferase Inhibitors. N-Myristoyltransferase (NMT) is an enzyme that catalyzes the transfer of myristic acid, from myristoyl-CoA to the N-terminal glycine's amine. This process is common for a variety of eukaryotic organisms.[26] Benzofuran compounds have been found to be selective C albicans NMT selective inhibitors. Since C albicans are organisms causing systemic fungal infections in immunocompromised patients, the compounds can be important drug candidates in AIDS therapy.[26]

In Figure 4 we illustrate the profiles describing IVE-PLS data elimination for the SOM and s- CoMSA methods, respectively. This allowed us to obtain final SOM-CoMSA ($q^2_{cv} = 0.84$) and s-CoMSA ($q^2_{cv} = 0.96$) models, which compare well with the Hasegawa SOM-CoMSA with a genetic algorithm (GA-SOM-CoMSA) model characterized by $q^2_{cv} = 0.81$.[26] Similar, to the models discussed above for asarones the sector method defined on the basis of the cubic grid appeared superior, providing slightly better final models. Figure 5 indicates the key surface sectors that are important for compound activity. Thus, a blue and red colored area indicates a positive contribution, which in this particular case increases the activity, while cyan and magenta sections generally decrease the activity. The Hasegawa model identifies for the electron-withdrawing substituents at the 2-, 3-, and 5-positions as the critical factors for activity. Generally, our model reveals a similar effect. Thus, the distribution of the electrostatic potential within the benzene ring decides the activity. For example, a large blue area determines the high activity of the compound **11b** in comparison to the low activity of **18b** (Figure 5).

From a theoretical point of view an interesting point of the Hasegawa series is the fact that individual compounds change only in the region of the aromatic ring. Thus, this region differs the most between compounds and should also be indicated as specific for the interaction in data elimination during QSAR analysis. In fact, this is true for the IVE-PLS models performed for the lower PLS latent variable numbers; however, the inclusion of a higher number of latent variables can also reveal possible interactions in the sectors that are common for all molecules. This is clearly illustrated by Figures 1 and 2 in the Supporting Information that analyze the areas of specific interactions as a function of the maximal number of latent variables that can be included in the model.

CONCLUSIONS

Shape analysis is a powerful tool in chemistry and drug design. In the current work, we compare the results of CoMFA and Comparative Molecular Surface Analysis

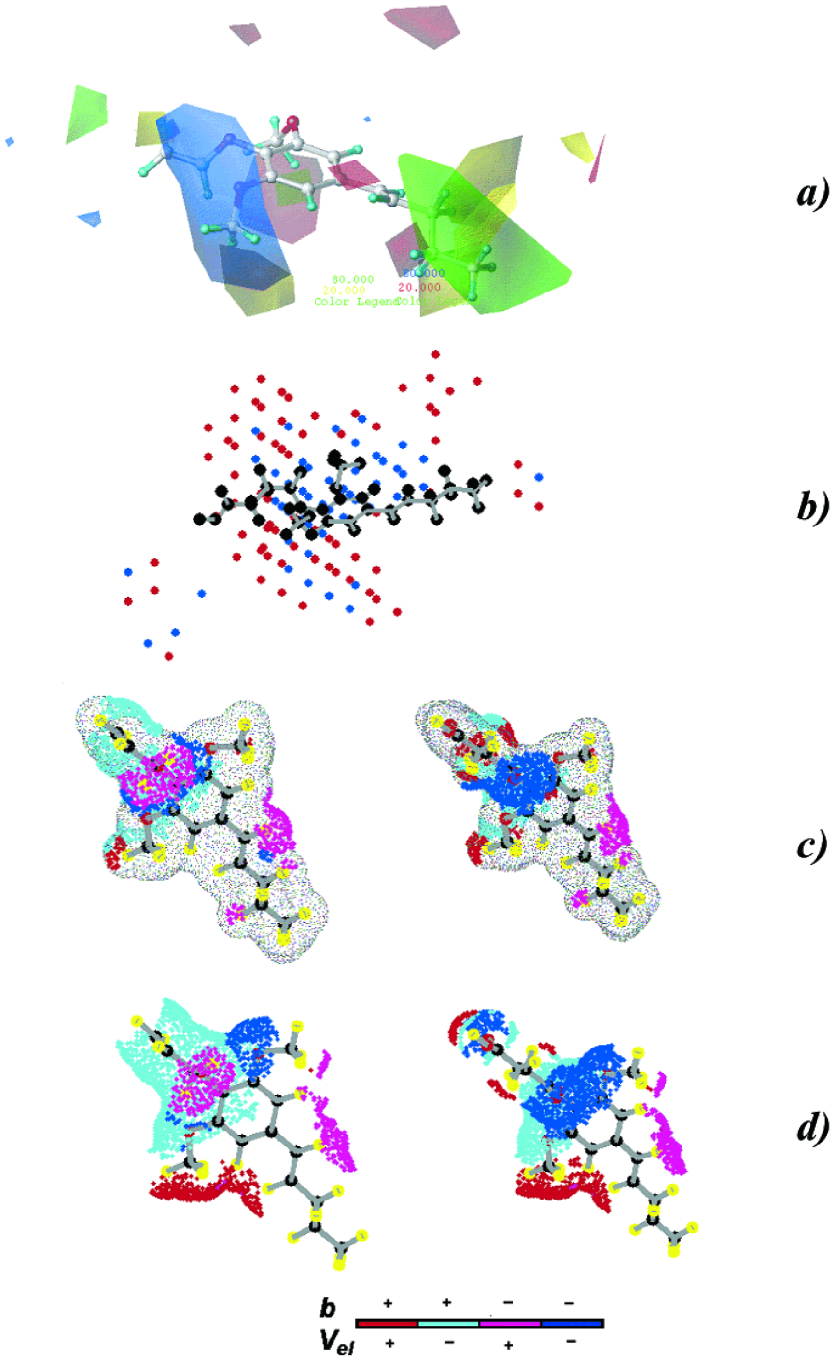


Figure 3. An illustration of the molecular areas that are key for the antipletelet asarones **a1**−**a40** activity, in a form of the traditional CoMFA profiles (a), CoMFA-IVE-PLS (b), SOM-CoMSA-IVE-PLS (c), and s-CoMSA-IVE-PLS (d). Compounds of the highest and lowest activity are shown for the CoMSA plots. Color codes indicate the following: for figure a − standard CoMFA coding; for figure b − blue − 50 points of the highest contribution to the model; red − next 100 points (out of 1650 points); figure c and d − a combination of the electrostatic potential sign and the sign of the b weight in the model, as shown by the colorbar: +/+ (red − decreases the activity), −/+ (cyan − increases the activity), +/− (magenta − increases the activity), −/− (blue − decreases the activity).

(CoMSA), the 3D QSAR method, for a series of hypolipidemic and antiplatelet asarones and antifungal N-myristoyltransferase inhibitors. In this publication we show that a sector CoMSA formalism enables an analysis of the biological activity that is more directly related to the molecular shape and individual molecular functionalities than the traditional uniform and directionless CoMFA field. Iterative Variable Elimination allowed us to identify the potential pharmacophoric sites. We modeled QSARs for both series and demonstrate that sector-based molecular descriptors give

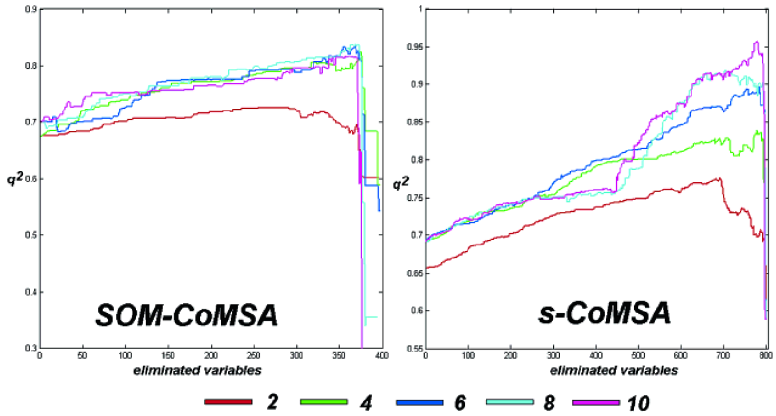


Figure 4. Profiles reporting the IVE-PLS modeling processes for the CoMFA and CoMSA, for the Hasegawa benzenefurans, details in text.

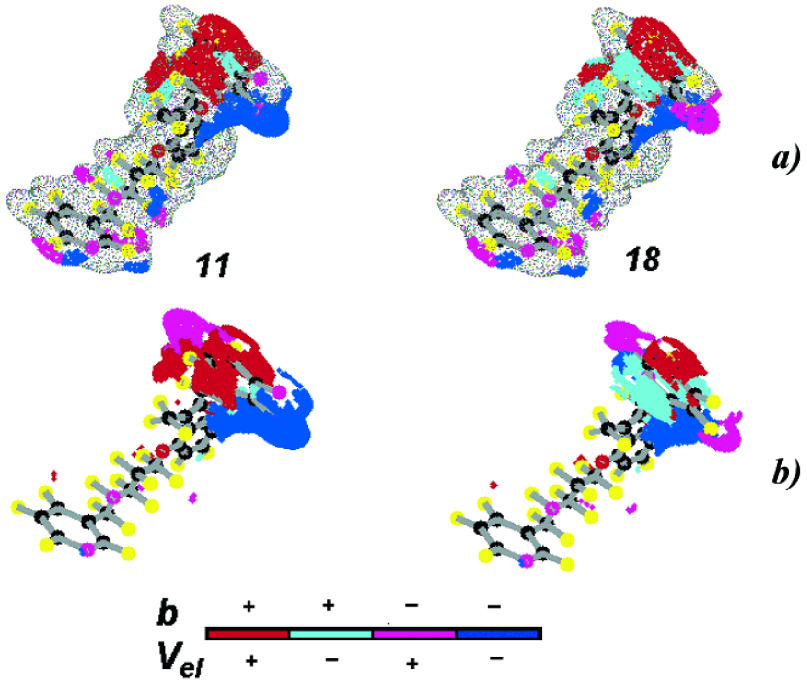


Figure 5. An illustration of the molecular areas that are key for the antifungal activity of compounds **b1**−**b29** (two compounds of the highest and lowest activity are shown): SOM-CoMSA-IVE-PLS (a) and s-CoMSA-IVE-PLS (b). Color codes indicate a combination of the electrostatic potential sign and the sign of the b weight in the model: +/+ (red − increases the activity), −/+ (cyan − decreases the activity), +/− (magenta − decreases the activity), −/− (blue − increases the activity), as shown by the colorbar.

very predictive models and allow one to generate a spatial interpretation of the QSAR models. In particular, we identified the central aromatic ring and carbonyl functions as the moieties determining the activity of the asarones series, while the pattern of substitution of the aromatic ring determines the activity of N-myristoyltransferase inhibitors.

ACKNOWLEDGMENT

The authors thank Professor Johann Gasteiger of the University of Erlangen-Nürnberg, BRD for facilitating access to the CORINA, PETRA, SURFACE, and KMAP programs. The financial support of the KBN Warsaw under grant nos. KBN 3T09 A01127 and PBZ 040 P04/08 is gratefully acknowledged. R.G. thanks Foundation for Polish Science for an individual grant.

Supporting Information Available: Illustrations of the molecular areas that are key for the antifungal activity as a function of the maximal number of the PLS latent variables included in the model for compounds **b1**−**b29** SOM-CoMSA-IVE-PLS (Figure 1) and s-CoMSA-IVE-PLS (Figure 2). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kolb, H. C.; Finn, M. G.; Sharpless, K. B. Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angew. Chem., Int. Ed.* **2001**, *40*, 2004−2021.

(2) Kubinyi, H. QSAR: Hansch Analysis and Related Approaches. In *Methods and Principles in Medicinal Chemistry*; Mannhold, R., Krogsgaard-Larsen, P., Timmerman, H., Eds.; VCH: Weinheim, 1993.
(3) Wermuth, C. G. The impact of QSAR and CADD methods on drug discovery. In *Rational Approaches to Drug Design − Proceedings of the 13th European Symposium on Quantitative Structure−Activity Relationships*; Holtje, H.-D., Sippl, W., Eds.; Prous: Barcelona, 2001; pp 3−20.
(4) Doweyko, A. M. 3D-QSAR illusions. *J. Comput.-Aided. Mol. Des.* **2004**, *18*, 587−96.
(5) Korhonen, S.-P.; Tuppurainen, K.; Laatikainen, R.; Peräkylä, M. FLUFF-BALL, A template-based grid-independent superposition and QSAR technique: Validation using a benchmark steroid data set. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1780−1793.
(6) Polanski, J.; Gieleciak, R.; Wyszomirski, M. Mapping dye pharmacophores by the comparative molecular surface analysis (CoMSA): application to heterocyclic monoazo dyes. *Dyes Pigm.* **2004**, *62*, 63−78.
(7) Polanski, J. Self-organizing neural networks for pharmacophore mapping. *Adv. Drug Deliv. Rev.* **2003**, *55*, 1149−1162.
(8) Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. Self-organizing neural networks for modeling robust 3D and 4D QSAR: Application to dihydrofolate reductase inhibitors. *Molecules* **2004**, *9*, 1148.
(9) Cramer, III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(10) Melville, J.; Hirst, J. D. On the stability of CoMFA models. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1294−1300.
(11) Jain, A. N.; Koile, K.; Chapman, D. Compass: Predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315−2327.
(12) Ivanciuc, O.; Ivanciuc, T.; Cabrol-Bass, D. 3D QSAR with CoRSA. Comparative receptor surface analysis. Application to calcium channel agonists. *Analusis* **2000**, *28*, 637−642.
(13) Polanski, J. Molecular shape analysis. In *Handbook of chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Verlag: Weinheim, 2003; pp 302−319.
(14) Hofbauer, C.; Aszodi, A. SH2 Binding site comparison: A new application of the SURFCOMP method. *J. Chem. Inf. Model.* **2005**, *45*, 414−421.
(15) Polanski, J.; Gieleciak, R.; Magdziarz, T. The grid formalism for the comparative molecular surface analysis: application to the CoMFA benchmark steroids, azo dyes and HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1423−1435.
(16) Anzali, S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J.; Teckentrup, A.; Wagener, M. The use of self-organizing neural networks in drug design. *Perspect. Drug Discovery Des.* **1998**, *9/10/11*, 273−299.
(17) Polanski, J.; Walczak, B. The comparative molecular surface analysis (CoMSA): a novel tool for molecular design. *Comput. Chem.* **2000**, *24*, 615−625.
(18) Polanski, J.; Gieleciak, R.; Bak, A. The comparative molecular surface analysis (CoMSA) − a nongrid 3D QSAR method by a coupled neural network and PLS system: Predicting pK_a values of benzoic and alkanoic acids. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 184−191.
(19) Polanski, J.; Gieleciak, R. The comparative molecular surface analysis (CoMSA) with modified uninformative variable elimination-PLS (UVE-PLS) method: application to the steroids binding the aromatase enzym. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 656−666.
(20) Polanski, J.; Gieleciak, R.; Wyszomirski, M. Comparative molecular surface analysis (CoMSA) for modeling dye-fiber affinities of the azo and antraquinone dyes. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1754−1762.
(21) Polanski, J.; Gieleciak, R. Comparative molecular surface analysis: a novel tool for drug design and molecular diversity studies. *Mol. Diversity* **2003**, *7*, 45−59.
(22) Polanski, J.; Gieleciak, R.; Bak, A. Probability issues in molecular design: Predictive and modeling ability in 3D-QSAR schemes. *Comb. Chem. High Throughput Screening* **2004**, *7*, 793−807.
(23) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D QSAR models using the 4D QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.
(24) Polanski, J.; Bak, A. Modeling steric and electronic effects in 3D and 4D-QSAR schemes: Predicting benzoic pK_a values and steroid CBG binding affinities. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2081−2092.
(25) Chilmonczyk, Z.; Siluk, D.; Kaliszan, R.; Lozowicka, B.; Poplawski, J.; Filipek, S. New chemical structures of hypolipidemic and antiplatelet activity. *Pure Appl. Chem.* **2001**, *73*, 1445−1458.
(26) Hasegawa, K.; Morikami, K.; Shiratori, Y.; Ohtsuka, T.; Aoki, Y.; Shimma, N. 3D-QSAR study of antifungal N-myristoyltransferase inhibitors by comparative molecular surface analysis. *Chemom. Intell. Lab. Syst.* **2003**, *69*, 51−59.
(27) Match3D program package, available from Professor J. Gasteiger, Computer-Chemie-Centrum, University Erlangen-Nurnberg, Germany. See: http://www2.ccc.uni-erlangen.de.
(28) Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatsu, K. New molecular surface-based 3D-QSAR method using Kohonen neural network and 3-Way PLS. *Comput. Chem.* **2002**, *26*, 583−589.
(29) Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatsu, K. Multi-way PLS modeling of structure−activity data by incorporating electrostatic and lipophilic potentials on molecular surface. *Comput. Biol. Chem.* **2003**, *27*, 381−386.
(30) MATLAB 5.0 program, available from: The Mathworks Inc., Natick, MA. http://www.mathworks.com.
(31) Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandeginste, B. M. V.; Sterna, C. Elimination of uninformative variables for multivariate calibration. *Anal. Chem.* **1996**, *68*, 3851−3858.
(32) Testa, B.; Purcell, W. P. A QSAR study of sulfonamide binding to carbonic anhydrase as test of steric models. *Eur. J. Med. Chem.* **1978**, *13*, 509−514.
(33) Eghdamian, B.; Ghose, K. Mode of action and adverse effects of lipid lowering drugs. *Drugs Today* **1998**, *34*, 943−956.
(34) Farmer, J. A.; Gotto, A., Jr. Current and future therapeutic approaches to hyperlipidemia. *Adv. Pharmacol.* **1996**, *35*, 79−114.
(35) Chamorro, G.; Garduno, L.; Sanchez, A.; Labarrios, F.; Salazar, M.; Martinez, E.; Diaz, F.; Tamariz, J. Hypolipemic activity of dimethoxy unconjugated propenyl side-chain analogues of alpha-asarone in mice. *Drug Dev. Res.* **1998**, *43*, 105−108.
(36) Labarrios, F.; Garduno, L.; Vidal, M.; Garcia, R.; Salazar, M.; Martinez, E.; Diaz, F.; Chamorro, G.; Tamariz, J. Synthesis and hypolipidaemic evaluation of a series of alpha-asarone analogues related to clofibrate in mice. *J. Pharm. Pharmacol.* **1999**, *51*, 1−7.
(37) Dandiya, P. C.; Sharma, J. D. Studies on Acorus calamus. V. Pharmacological actions of asarone and beta-asarone on central nervous system. *Indian J. Med. Res.* **1962**, *50*, 46−60.
(38) Dandiya, P. C.; Menon, M. K. Effects of asarone and beta-asarone on conditioned responses, fighting behaviour and convulsions. *Br. Pharm. Chemother.* **1963**, *20*, 436−442.
(39) Belova, L.; Alibekov, S.; Baginskaya, A.; Sokolov, S.; Pokrovskaya, G.; Stikhin, V.; Trumpe, T.; Gorodnyuk, T. Asarone and its biological properties. *Farmak. Toksikol.* **1985**, *48*, 17−20.
(40) Salazar, M.; Salazar, S.; Ulloa, V.; Mendoza, T.; Pages, N.; Chamorro, G. Teratogenic action of alpha-asarone in the mouse. *J. Toxicol. Clin. Exp.* **1992**, *12*, 149−154.
(41) Filipek, S.; Lozowicka, B. Alpha-asarone congeners as hypolipidemic agents. Pseudoreceptor versus minireceptor modeling. *Acta Pol. Pharm.* **2000**, *57*, 106−109.
(42) Cruz, M.; Salazar, M.; Garciafigueroa, Y.; Hernandez, D.; Diaz, F.; Chamorro, G.; Tamariz, J. Hypolipidemic activity of new phenoxyacetic derivatives related to alpha asarone with minimal pharmacophore features *Drug. Dev. Res.* **2003**, *60*, 186−190.
(43) Magdziarz, T.; Lozowicka, B.; Gieleciak, R.; Bak, A.; Polanski, J.; Chilmonczyk, Z. The 3D QSAR study of hypolipidemic asarones by comparative molecular surface analysis. *Bioorg. Med. Chem.* submitted.

CI0501488